Predicitive Model Selection for Tracheostomy or Death in Neonates with Severe Bronchopulmonary Dysplasia (sBPD)

Angel Zheng

Nov 2023 for PHP2550

Abstract

Infants diagnosed with severe bronchopulmonary dysplasia (sBPD) often undergo tracheostomy prior to discharge to facilitate their daily living and maintain vital functions. In this report, we construct a predictive model for the outcomes of tracheostomy or death in infants with sBPD. Two distinct model development methods, namely LASSO and Best Subset, were employed, resulting in four versions of the model with varying variable inclusions. Each method computes two models, with and without interactions. To enhance the robustness of our findings, the model development process incorporated multiple imputation and cross-validation techniques. This involved averaging set of variable coefficients with lowest error across imputed datasets to derive the final model. As a result, our comparative analysis revealed that the Lasso method with interactions emerged as the most effective approach for predicting tracheostomy or death outcomes in infants with sBPD. The proposed model holds promise for informing clinical decision-making and improving the overall care and prognosis for this patient population.

1 Introduction

[1]Bronchopulmonary Dysplasia (BPD) stands as one of the prevailing complications affecting premature infants, with classifications ranging from mild to moderate and severe. This condition arises in tandem with the prematurity of birth, a scenario where infants are delivered before reaching the expected gestational maturity, leading to structural damage in the lungs. Annually, severe Bronchopulmonary Dysplasia (sBPD) impacts over 10,000 neonates, necessitating the dependence of afflicted infants on ventilators to bolster lung function. Remarkably, 75% of these infants are discharged from medical facilities with ventilator support, facilitated by tracheostomy, ensuring ongoing assistance for daily living and the preservation of vital functions in the future.

Despite tracheostomy being a commonplace surgical intervention offering support to patients with sBPD, it is not without its inherent risks, notably an increased susceptibility to death and infection. Consequently, within clinical settings, the imperative arises to discriminate among sBPD patients, determining those who genuinely require tracheostomy intervention. Recognizing this critical need, the focus of this report is the development of statistical models employing LASSO (Least Absolute Shrinkage and Selection Operator), Best Subset, and Forward Selection. Augmented by multiple imputation and cross-validation techniques, the models aim to address the challenge of predicting whether a patient with sBPD is likely to necessitate a tracheostomy.

Moreover, it is important to highlight the current absence of an established model for predicting the necessity of tracheostomy in sBPD cases. As a result, the decision to proceed with tracheostomy often relies heavily on the subjective judgment of clinicians, underscoring a critical gap in evidence-based decision-making. This report endeavors to fill this gap by developing and evaluating predictive models that offer more objective insights into the likelihood of tracheostomy in infants with severe bronchopulmonary dysplasia. The model selection process, encompassing variable selection, is geared towards identifying the most influential predictors in infants with sBPD for tracheostomy outcomes.

2 Study Population

[2]Study participants were selected from the BPD Collaborative Registry, a multi-center consortium comprising interdisciplinary BPD programs situated in the United States and Sweden. The participants consisted of infants diagnosed with severe Bronchopulmonary Dysplasia (sBPD) from nine distinct centers. Within the registry, a comprehensive dataset of standard demographic and clinical information was systematically collected at four pivotal time points: birth, 36 weeks postmenstrual age (PMA), 44 weeks PMA, and discharge.

To depict the study population comprehensively, two summary tables were generated, each incorporating variables collected by the registry across multiple time points for individual patients but stratified differently. The initial table is categorized by the composite outcome, namely the occurrence of tracheostomy or death before discharge. Notably, there is a larger cohort of patients with no tracheostomy or death (N=811) compared to those with the composite outcome (N=183). While minimal disparities are observed in the birth variables, discernible differences emerge in the proportions of patients with elevated ventilation support levels and medication usage at both 36 weeks and 44 weeks among those with the positive outcome. This emphasizes the potential association of ventilation support levels and medication with the likelihood of tracheostomy or death in severe Bronchopulmonary Dysplasia (sBPD) patients. Furthermore, patients with a positive outcome exhibit a higher hospital discharge gestational age, suggesting a potential correlation between gestational age at discharge and the composite outcome.

Table 1: Summary Statistics by Outcome (Tracheostormy/Death vs. None)

Characteristic	No Tracheostomy nor Death(N = 811)	Tracheostomy or Death (N = 183
center		
1	25 (3.1%)	31 (17%)
2	545 (68%)	84 (47%)
3	55 (6.8%)	2 (1.1%)
4	47 (5.8%)	12 (6.7%)
5	33 (4.1%)	7 (3.9%)
7	31 (3.9%)	1 (0.6%)
12	28 (3.5%)	41 (23%)
16	37 (4.6%)	1 (0.6%)
20	4 (0.5%)	0 (0%)
Birth weight (g)	760 (610, 950)	670 (540, 835)
Obstetrical gestational age	25 (24, 27)	25 (24, 27)
Birth length (cm)	32 (30, 35)	31 (29, 34)
Birth head circumference (cm)	23.00 (21.50, 25.00)	22.00 (21.00, 24.00)
Delivery method (1=vaginal, 2=cesarean)		
1	245 (30%)	39 (21%)
2	564 (70%)	143 (79%)
Prenatal Corticosteroids	679 (86%)	154 (92%)
Complete Prenatal Steroids	499 (65%)	109 (70%)
Maternal Chorioamnionitis	132 (17%)	28 (17%)
gender		
Female	334 (41%)	73 (40%)
Male	473 (59%)	110 (60%)
Small for Gestational Age		
Not SGA	658 (82%)	118 (66%)
SGA	142 (18%)	61 (34%)
Surfactant in first 72 hrs	374 (81%)	87 (88%)
Weight at 36 Weeks	2,150 (1,880, 2,408)	1,997 (1,694, 2,260)
Ventilation Support Level at 36 Weeks		
0	109 (14%)	7 (4.3%)
1	553 (69%)	36 (22%)
2	140 (17%)	119 (73%)
Fraction of Inspired Oxygen at 36 Weeks	0.29 (0.23, 0.35)	0.45 (0.34, 0.60)
Peak Inspiratory Pressure at 36 Weeks	0 (0, 0)	14 (2, 24)

Table 1: Summary Statistics by Outcome (Tracheostormy/Death vs. None) *(continued)*

Characteristic	No Tracheostomy nor Death(N = 811)	Tracheostomy or Death (N = 183
Positive and Exploratory Pressure at 36 Weeks	7 (6, 8)	8 (6, 9)
Medication for Pulmonary Hypertension at 36 Weeks		
0	770 (96%)	129 (80%)
1	32 (4.0%)	33 (20%)
Weight at 44 Weeks	3,765 (3,299, 4,143)	3,555 (3,070, 3,995)
Ventilation Support Level at 44 Weeks		
0	261 (60%)	8 (6.0%)
1	124 (28%)	22 (16%)
2	53 (12%)	104 (78%)
Fraction of Inspired Oxygen at 44 Weeks	0.28 (0.25, 0.32)	0.40 (0.30, 0.60)
Peak Inspiratory Pressure at 44 Weeks	0 (0, 0)	19 (10, 37)
Positive and Exploratory Pressure at 44 Weeks	0 (0, 8)	8 (8, 10)
Medication for Pulmonary Hypertension at 44 Weeks		
0	405 (92%)	68 (51%)
1	33 (7.5%)	66 (49%)
Hospital Discharge Gestational Age	45 (42, 52)	64 (50, 91)

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

The earlier table displayed the proportion of patients with outcomes across nine different centers. Notably, Centers 1 and 12 had more patients with tracheostomy or death, while most centers had more severe Bronchopulmonary Dysplasia (sBPD) patients without tracheostomy before discharge. This suggests potential baseline differences among the centers, as highlighted in Table 2, which breaks down the data by center.

In Table 2, it's clear that Centers 1 and 12 had patients with a higher level of ventilation support and more medication at both 36 and 44 weeks postmenstrual age (PMA). Other variables did not show significant differences. This suggests that these two centers might be different from the others in terms of the severity of patients they handle.

Table 2: Summary Statistics by Center

Characteristic	**1**, N = 56	**2**, N = 630	**3**, N = 57	**4**, N = 60	**5**, N = 40	**7**, N = 32	**12**, N = 69	**16**, N = 38	**20**, N = 4	**p-value**
Birth weight (g)	649 (539, 790)	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)	<0.001
Obstetrical gestational age	25 (24, 27)	26 (24, 27)	26 (24, 27)	25 (24, 27)	24 (23, 25)	25 (23, 27)	26 (25, 27)	26 (24, 28)	26 (25, 27)	<0.001
Birth length (cm)	31 (29, 33)	32 (30, 35)	32 (31, 34)	33 (31, 35)	29 (28, 31)	32 (29, 36)	33 (31, 34)	33 (31, 37)	34 (29, 38)	<0.001
Birth head circumference (cm)	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)	24.90 (22.94, 25.98)	<0.001
Delivery method (1=vaginal, 2=cesarean)										
1	15 (27%)	177 (28%)	17 (30%)	18 (30%)	14 (35%)	10 (31%)	18 (27%)	14 (37%)	1 (25%)	
2	41 (73%)	453 (72%)	40 (70%)	42 (70%)	26 (65%)	22 (69%)	49 (73%)	24 (63%)	3 (75%)	
Prenatal Corticosteroids	47 (90%)	544 (86%)	47 (87%)	47 (80%)	37 (93%)	26 (87%)	41 (89%)	33 (89%)	3 (75%)	
Complete Prenatal Steroids	29 (63%)	415 (68%)	41 (76%)	25 (45%)	27 (68%)	15 (56%)	26 (59%)	26 (70%)	1 (33%)	
Maternal Chorioamnionitis	14 (47%)	105 (17%)	4 (12%)	8 (14%)	14 (36%)	3 (9.7%)	6 (8.7%)	2 (6.1%)	1 (33%)	
gender										
Female	22 (39%)	249 (40%)	22 (39%)	28 (47%)	17 (43%)	16 (50%)	28 (41%)	20 (53%)	1 (25%)	
Male	34 (61%)	379 (60%)	34 (61%)	32 (53%)	23 (58%)	16 (50%)	41 (59%)	18 (47%)	3 (75%)	
Small for Gestational Age										
Not SGA	34 (62%)	503 (81%)	41 (76%)	54 (92%)	32 (80%)	24 (75%)	52 (75%)	31 (82%)	2 (50%)	
SGA	21 (38%)	117 (19%)	13 (24%)	5 (8.5%)	8 (20%)	8 (25%)	17 (25%)	7 (18%)	2 (50%)	
Surfactant in first 72 hrs	31 (91%)	265 (79%)	54 (96%)	11 (69%)	33 (94%)	3 (50%)	48 (84%)	9 (56%)	2 (100%)	
Weight at 36 Weeks	2,100 (1,808, 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)	0.004
Ventilation Support Level at 36 Weeks										
0	7 (13%)	50 (8.1%)	5 (8.9%)	8 (13%)	0 (0%)	22 (69%)	1 (2.0%)	22 (58%)	1 (25%)	
1	20 (36%)	425 (68%)	35 (63%)	34 (57%)	31 (78%)	8 (25%)	17 (34%)	14 (37%)	2 (50%)	
2	29 (52%)	146 (24%)	16 (29%)	18 (30%)	9 (23%)	2 (6.3%)	32 (64%)	2 (5.3%)	1 (25%)	
Fraction of Inspired Oxygen at 36 Weeks	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)	<0.001
Peak Inspiratory Pressure at 36 Weeks	3 (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)	<0.001
Positive and Exploratory Pressure at 36 Weeks	8 (7, 10)	7 (6, 8)	8 (7, 10)	6 (6, 7)	9 (8, 10)	0 (0, 5)	6 (6, 7)	0 (0, 8)	6 (4, 6)	<0.001
Medication for Pulmonary Hypertension at 36 Weeks										
0	42 (75%)	596 (96%)	53 (95%)	49 (82%)	37 (93%)	30 (94%)	46 (92%)	34 (89%)	4 (100%)	
1	14 (25%)	25 (4.0%)	3 (5.4%)	11 (18%)	3 (7.5%)	2 (6.3%)	4 (8.0%)	4 (11%)	0 (0%)	
Weight at 44 Weeks	3,695 (3,154, 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)	0.018
Ventilation Support Level at 44 Weeks										
0	9 (17%)	198 (51%)	12 (60%)	0 (NA%)	19 (61%)	10 (83%)	12 (26%)	5 (100%)	2 (50%)	
1	15 (29%)	97 (25%)	7 (35%)	0 (NA%)	9 (29%)	0 (0%)	13 (28%)	0 (0%)	1 (25%)	
2	28 (54%)	96 (25%)	1 (5.0%)	0 (NA%)	3 (9.7%)	2 (17%)	22 (47%)	0 (0%)	1 (25%)	
Fraction of Inspired Oxygen at 44 Weeks	0.31 (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)	0.36 (0.29, 0.42)	0.040
Peak Inspiratory Pressure at 44 Weeks	12 (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)	<0.001
Positive and Exploratory Pressure at 44 Weeks	9 (6, 12)	0 (0, 8)	0 (0, 6)	NA (NA, NA)	0 (0, 8)	0 (0, 0)	6 (0, 8)	0 (0, 0)	4 (0, 7)	<0.001
Medication for Pulmonary Hypertension at 44 Weeks										
0	25 (48%)	350 (90%)	19 (95%)	0 (NA%)	26 (84%)	8 (67%)	30 (64%)	4 (80%)	4 (100%)	
1	27 (52%)	41 (10%)	1 (5.0%)	0 (NA%)	5 (16%)	4 (33%)	17 (36%)	1 (20%)	0 (0%)	
Hospital Discharge Gestational Age	63 (61, 64)	47 (42, 55)	44 (41, 45)	NA (NA, NA)	52 (48, 54)	43 (39, 47)	51 (47, 59)	40 (39, 43)	52 (44, 63)	<0.001
Trach_or_Death										
0	25 (45%)	545 (87%)	55 (96%)	47 (80%)	33 (83%)	31 (97%)	28 (41%)	37 (97%)	4 (100%)	
1	31 (55%)	84 (13%)	2 (3.5%)	12 (20%)	7 (18%)	1 (3.1%)	41 (59%)	1 (2.6%)	0 (0%)	

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

[2] Kruskal-Wallis rank sum test

3 Method

The code for the procedures outlined in this section is available on GitHub and is executed using the statistical software R (version 4.3.1).

Multiple Imputation and Test Split

Multiple imputation creates multiple copies of the dataset with different imputed values and does identical analysis on all of the copies. There are three phases that are in multiple imputation tool: the imputation phase, which repetitively applies algorithms to generate fill-in entries for the missing; the analysis phase, which utilizes statistical analysis as if there weren't missingness; lastly the pooling phase, which averages out the multiple sets of parameters for each copy.

When data are Missing Completely at Random (MCAR, the probability of missing data on a variable X is not related to other measured variables or X itself) or Missing at Random (MAR: the probability of missing data on X is related to other measured variables, but not to the underlying values of X itself), multiple imputation is an useful tool to handle missingness and therefore maintain original study population even with some missed observations.

And because the study population now has large enough subject, the subjects are split into "train" and "test" subsets in order to test model performance. This split is randomized by software, and the proportion is set manually to 70% train and 30% test.

Cross-Validated Model Selection

Cross-validation is used in combination with LASSO and Best subset methods to develop four versions of the predictive model. Cross-validation is a statistical technique employed to assess the performance and generalizability of a predictive model. It involves partitioning the dataset into multiple subsets, commonly referred to as folds, with one subset reserved for testing the model and the others used for training. This process is iteratively repeated, each time using a different fold for testing. The results from each iteration are then averaged to provide a robust evaluation of the model's performance, minimizing the risk of overfitting or underfitting to a specific subset of data.

This report employs LASSO and Best Subset methods for model selection, coupled with cross-validation. These techniques, in conjunction with cross-validation, derive a set of coefficients, leading to the elimination of some variables from the predictor pool due to insignificance. Each method employs distinct algorithms in the variable selection process, and the resulting models' performances are subsequently compared. The model development process utilizes the "train" sets, while the model evaluation process relies on the "test" set.

$$\text{LASSO method minimizes } \frac{1}{n}\Sigma_{i=1}^{n}\left(y_i - \beta x_i\right)^2 + \lambda\Sigma_{j=1}^{p}\left|\beta_j\right|$$

$$\text{Best Subset method minimizes } \frac{1}{n}\Sigma_{i=1}^{n}\left(y_i - \beta x_i\right)^2 + \lambda\Sigma_{j=1}^{p}1\cdot\left(\beta_j \neq 0\right)$$

Model Evaluation

The result models are evaluated using Sensitivity, Specificity, Brier Score, and AUC.

4 Results

Correlation and Missingness

In this figure, the correlation between variables are plotted. Variables at birth are inter-correlated with each other which is reasonable because the variables are weight, length, and head circumference which are all related to an infant's size. Obstetrical gestational age is positively correlated to the size of an infant, that if the gestational age is low the size is small. Another observable correlation is between weight at 36 weeks and weight at 44 weeks. No collinearity is found.

Looking at the correlations, there could be possible interactions between the respiratory variables within same timepoint, and would be added to the predictor pool in the next step.

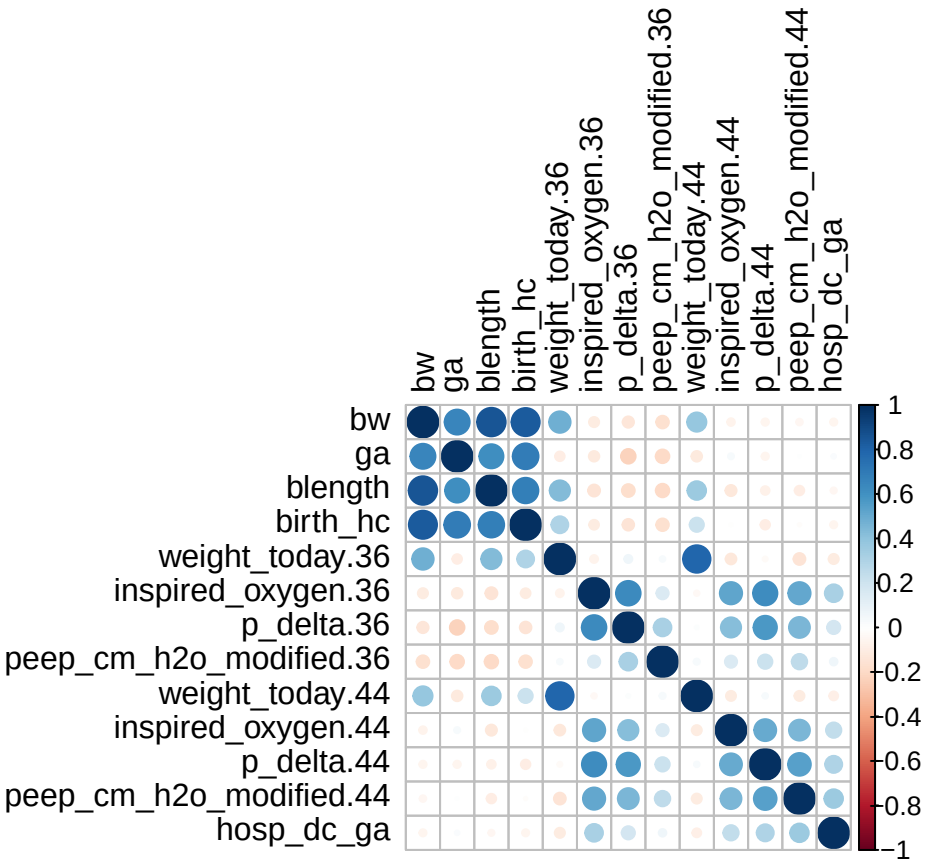

Figure 1: Correlation Plot for Variables

EDA for missingness

The missingness summary table shows that this dataset has missingness for all variables at 44 weeks as well as the variable `any_surf`. It is observable that most missingness is at 44 weeks level, and is due to the early discharge age. If a patient discharged before 44 weeks PMA, their data is noted as NA. This missingness can be explained by observable factor, therefore the assumption holds for implementing multiple imputation.

Table 3: Missingness Summary

Variable Name	Number of Missing	Percentage % of Missing
Fraction of Inspired Oxygen at 44 Weeks	448	44.9799197
Peak Inspiratory Pressure at 44 Weeks	448	44.9799197
Weight at 44 Weeks	446	44.7791165
Positive and Exploratory Pressure at 44 Weeks	446	44.7791165
Surfactant in first 72 hrs	433	43.4738956
Ventilation Support Level at 44 Weeks	424	42.5702811
Medication for Pulmonary Hypertension at 44 Weeks	424	42.5702811
Complete Prenatal Steroids	128	12.8514056
Peak Inspiratory Pressure at 36 Weeks	124	12.4497992
Hospital Discharge Gestational Age	117	11.7469880
Positive and Exploratory Pressure at 36 Weeks	92	9.2369478
Weight at 36 Weeks	92	9.2369478
Fraction of Inspired Oxygen at 36 Weeks	78	7.8313253
Birth length (cm)	77	7.7309237
Birth head circumference (cm)	71	7.1285141
Maternal Chorioamnionitis	62	6.2248996
Maternal Ethnicity	57	5.7228916
Prenatal Corticosteroids	35	3.5140562
Ventilation Support Level at 36 Weeks	30	3.0120482
Medication for Pulmonary Hypertension at 36 Weeks	30	3.0120482
Small for Gestational Age	15	1.5060241
Center	10	1.0040161
Gender	4	0.4016064
Delivery method	3	0.3012048
Outcome - Tracheostomy or Death	2	0.2008032

Variable and Model Selection

The two methods compute different sets of averaged coefficients, and are summarized in the below table. Exclusion occurrence means the count of how many times this variable being excluded in the model output (i.e. if exclusion occurrence = 5, this variable never appeared when the model selection is performed on the 5 imputed train set respectively). Variables highlighted have low count of exclusion in both methods, meaning that they are chosen as predictors for both models.

Table 4: Predictors at Birth level

	occurrence_zeros_lasso	coefs_lasso	occurrence_zeros_bs	coefs_bs
bw	3	0.00007	5	0.00000
del_method2	0	0.59559	0	0.95982
prenat_sterYes	2	0.17167	5	0.00000
com_prenat_sterYes	0	0.43304	2	0.44738
mat_chorioYes	3	0.10667	5	0.00000
genderMale	3	0.05518	5	0.00000
sgaSGA	3	0.00127	5	0.00000
any_surfYes	0	0.05757	3	-0.28321

Table 5: Predictors at 36 wk PMA level

	occurrence_zeros_lasso	coefs_lasso	occurrence_zeros_bs	coefs_bs
weight_today.36	5	0.00000	4	0.00025
ventilation_support_level.361	0	-0.39202	5	0.00000
ventilation_support_level.362	0	1.39655	0	1.99057
inspired_oxygen.36	0	1.44251	2	1.35199
p_delta.36	4	0.00094	5	0.00000
peep_cm_h2o_modified.36	4	0.00693	4	0.01360
med_ph.361	4	0.01229	4	-0.19513

Table 6: Predictors at 44 wk PMA level

	occurrence_zeros_lasso	coefs_lasso	occurrence_zeros_bs	coefs_bs
weight_today.44	0	-0.00061	0	-0.00081
ventilation_support_level_modified.441	3	-0.52550	5	0.00000
ventilation_support_level_modified.442	1	0.98464	0	2.41599
inspired_oxygen.44	0	1.39885	2	1.52410
p_delta.44	4	0.00091	4	-0.00753
peep_cm_h2o_modified.44	0	0.15329	3	0.06499
med_ph.441	1	0.41807	4	0.10869

Table 7: Interaction-Term Predictors

	occurrence_zeros_lasso	coefs_lasso	occurrence_zeros_bs	coefs_bs
ventilation_support_level.361:med_ph.361	0	0.83195	1	1.27270
ventilation_support_level.362:med_ph.361	4	-0.07371	5	0.00000
ventilation_support_level_modified.441:p_delta.44	0	0.09678	0	0.16171
ventilation_support_level_modified.441:med_ph.441	1	1.22199	2	1.79672
ventilation_support_level_modified.442:med_ph.441	2	-0.72757	3	-0.95102
inspired_oxygen.44:med_ph.441	1	1.64224	3	2.72390

The coefficients are then multiplied with the corresponding variable from test dataset, and the probability of outcome is calculated as

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \ldots + \beta_p x_p)}}$$

Discrimination and Calibration

```
## [1] 0.8517241 0.7585366 0.1059817 0.8776339
```

```
## [1] 0.8793103 0.7276423 0.1050427 0.8852033
```

```
## [1] 0.7862069 0.7894309 0.1075775 0.8676983
```

```
## [1] 0.8344828 0.7593496 0.1071936 0.8808074
```

Table 8: Model Evaluation Metrics

	Lasso	Lasso with Interactions	Best Subset	Best Subset with Interactions
Sensitivity	0.8517241	0.8793103	0.7862069	0.8344828
Specificity	0.7585366	0.7276423	0.7894309	0.7593496
Brier Score	0.1059817	0.1050427	0.1075775	0.1071936
AUC	0.8776339	0.8852033	0.8676983	0.8808074

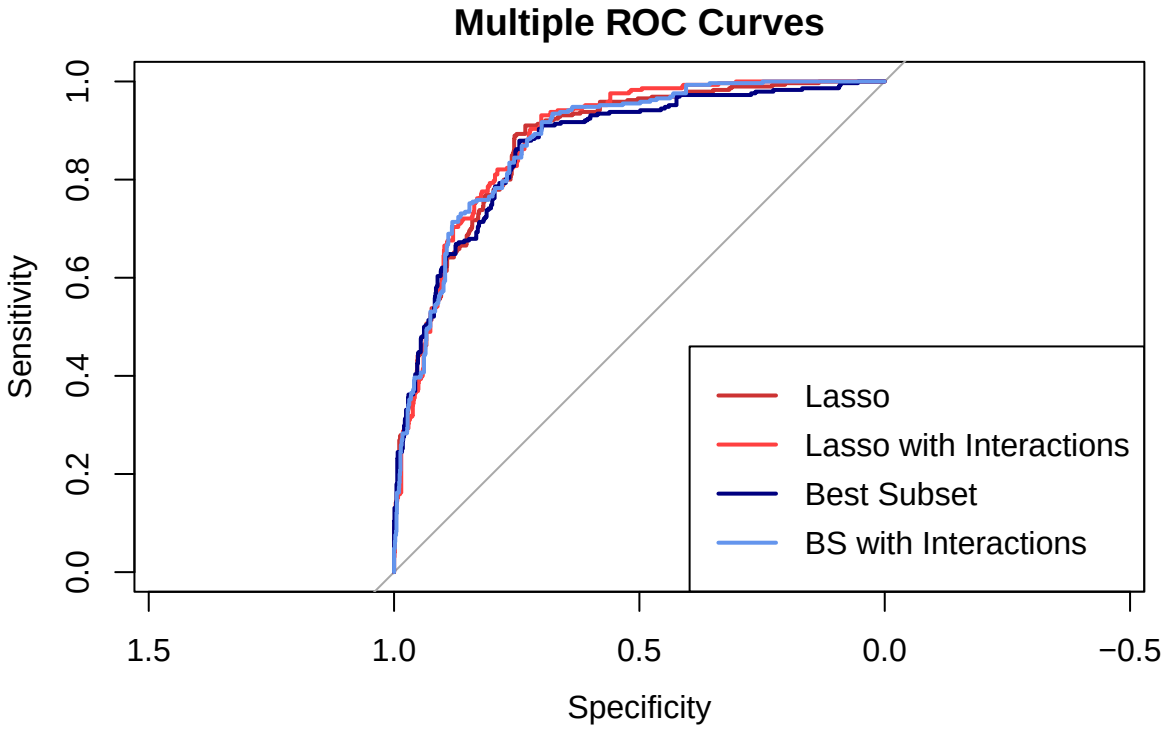

Figure 2: ROC Curves for Three Models

5 Discussion

The four models in our analysis compute distinct sets of coefficients, with each set derived through five iterations of model selection on the training data, subsequently averaging the outcomes. Both the Lasso and Best Subset methods exhibit similar performance characteristics while Best Subset being slightly more aggresive in variable selection than Lasso, specifically within this dataset. By fitting the model with and without interaction terms, by evaluation metrics it is seen that both methods perform better (lower Brier score, higher AUC) with inclusion of interaction terms. Then, Lasso model with interactions is a more accurate fit than Best Subset model with interactions, as seen in the two calibration plots.

Notably, while the center variable is retained in both the Lasso and Best Subset approaches, its presence may limit the generalizability of the final model. To assess the impact of excluding the center variable on overall model performance, a test is conducted. The results reveal that when dropping the center variable while retaining the other chosen variables from the cross-validated Lasso, the model performs slightly worse than the original. This underscores the nuanced influence of `center` variables on the overall predictive capacity of the model.

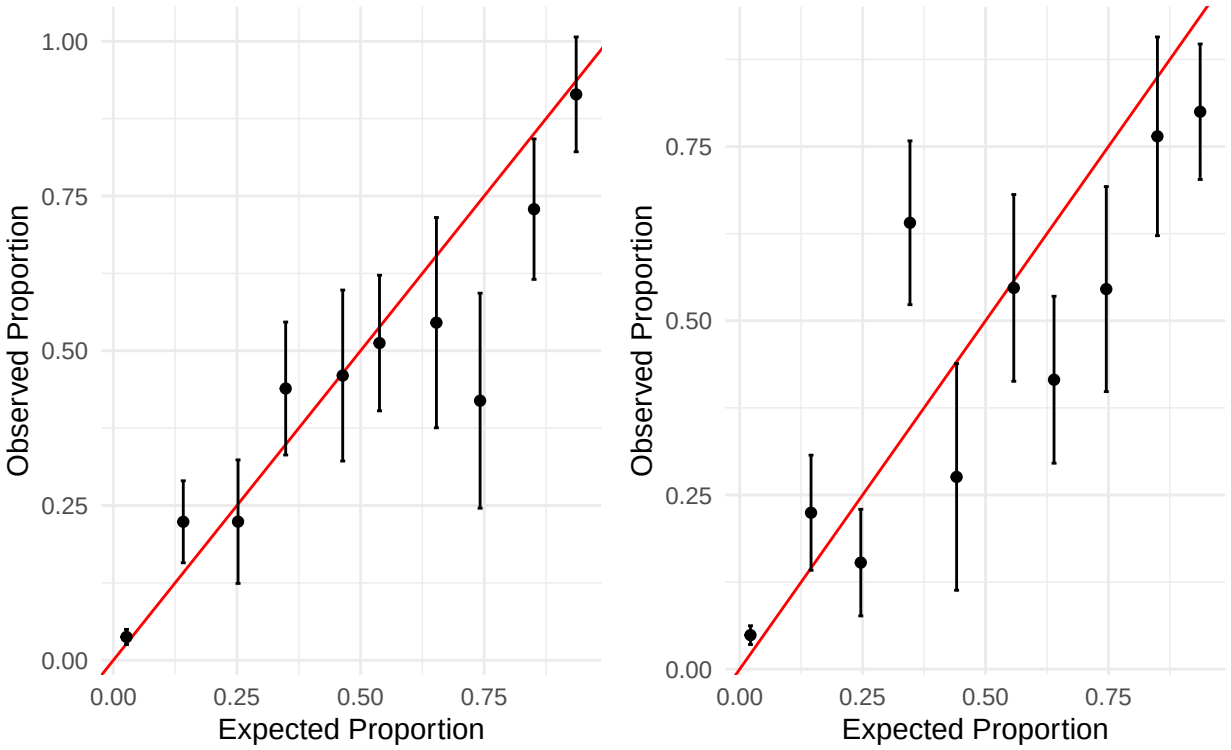

Figure 3: Calibration Plots for Lasso and Best Subset Models

Strength and Limitation

Despite achieving a low Brier score and a high AUC value, indicating strong predictive performance, the final model derived from cross-validated Lasso exhibits a potential limitation in its generalizability to data from centers not present in the original dataset. This suggests a need for caution when applying the model beyond the observed centers. Nevertheless, the model, which retains the majority of variables, demonstrates relatively high accuracy within the known centers.

Another limitation of the model stems from missing data, particularly at 44 weeks and across various centers. Multiple imputation assumes Missing at Random (MAR) conditions, yet the missingness, especially at 44 weeks, introduces uncertainty into whether the imputed dataset accurately reflects the broader population.

6 Conclusion

In summary, among the Lasso and Best Subset Selection methods, the Lasso method emerges as the most effective in constructing a predictive model for Tracheostomy or Death in Neonates with Severe Bronchopulmonary Dysplasia (sBPD). Futhermore, by inputting pre-selected interaction terms into Lasso allows selection on significant interactions. Result showed that the model with interactions is in fact a better fit than without interactions. Some variables are selected by both Lasso and Best Subset model, which is a sign that those variables are particularly important. Coefficients demonstrate cesarean delivery and invasive ventilation support at both 36 and 44 weeks PMA are important risk predictors for tracheostomy. Therefore, by utilizing this model in clinical setting, clinicians can predict the suitability of tracheostomy based on patient information at birth, 36 weeks, and 44 weeks.

6 Reference

[1] U.S. Department of Health and Human Services. (n.d.). Bronchopulmonary dysplasia (BPD). National Heart Lung and Blood Institute. https://www.nhlbi.nih.gov/health/bronchopulmonary-dysplasia

[2] Data information given by Dr. Robin Mckinney

7 Appendix: All code for this report

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning=FALSE)
library(dplyr)
library(tidyverse)
library(ggmice)
library(mice)
library(gtsummary)
library(glmnet)
library(bestglm)
library(kableExtra)
library(leaps)
library(LOLearn)
library(Matrix)
library(nestfs)
library(pROC)
library(ggpubr)
library(naniar)
library(corrplot)

# load dataset
sBPD <- read.csv("C:/ANGEL/Brown 22-24/23Fall/PHP2550_pda/datasets/project2.csv")

# correct a miscoded subject from dataset
#sBPD[which(sBPD$center == 21),]
sBPD[810,]$record_id <- 1000001
sBPD[810,]$center <- 1

# composite outcome variable
sBPD <- sBPD %>%
  arrange(record_id) %>%
  select(-mat_race) %>%
  mutate(Trach_or_Death = if_else(Trach == 1 | Death == "Yes", 1, 0)) %>%
  select(-c(Trach, Death)) %>%
  unique()

# factor variables
sBPD[,c(2:3, 8:14, 16, 20, 22, 26, 28)] <- lapply(sBPD[,c(2:3, 8:14, 16, 20, 22, 26, 28)], as.factor)

# fill in some complete prenatal steroids for those not given prenatal corticosteroids
sBPD$com_prenat_ster[sBPD$prenat_ster == "No"] <- "No"
# variable names for table summary
var_label <- c(bw ~ "Birth weight (g)",
               ga ~ "Obstetrical gestational age",
               blength ~ "Birth length (cm)",
               birth_hc ~ "Birth head circumference (cm)",
               del_method ~ "Delivery method (1=vaginal, 2=cesarean)",
               prenat_ster ~ "Prenatal Corticosteroids",
               com_prenat_ster ~ "Complete Prenatal Steroids",
               mat_chorio ~ "Maternal Chorioamnionitis",
               sga ~ "Small for Gestational Age",
               any_surf ~ "Surfactant in first 72 hrs",
               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 at 36 Weeks",
                peep_cm_h2o_modified.36 ~ "Positive and Exploratory Pressure 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 at 44 Weeks",
                peep_cm_h2o_modified.44 ~ "Positive and Exploratory Pressure at 44 Weeks",
                med_ph.44 ~ "Medication for Pulmonary Hypertension at 44 Weeks",
                hosp_dc_ga ~ "Hospital Discharge Gestational Age")

# summary table by outcome
sBPD %>%
  tbl_summary(include = -c(record_id, mat_ethn),
              by = Trach_or_Death,
              missing = "no",
              label = var_label)  %>%
  modify_header(list(stat_1 ~ "No Tracheostomy nor Death(N = 811)",
                     stat_2 ~ "Tracheostomy or Death (N = 183")) %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Summary Statistics by Outcome (Tracheostormy/Death vs. None)",
                 longtable = TRUE) %>%
  kableExtra::kable_styling(font_size = 8, latex_options = c("repeat_header", "HOLD_position", "scale_d

# summary table by center
sBPD %>%
  tbl_summary(include = -c(record_id, mat_ethn),
              by = center,
              missing = "no",
              label = var_label)%>%
  add_p() %>%
  as_kable_extra(booktabs = TRUE,
    caption = "Summary Statistics by Center") %>%
  kableExtra::kable_styling(font_size = 8,
    latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  landscape()

# random vector to indicate test or train
set.seed(1)
ignore <- sample(c(TRUE, FALSE), size = nrow(sBPD), replace = TRUE, prob = c(0.3, 0.7))

# multiple imputation
sBPD_mice_out <- mice(sBPD[,-1], m = 5, seed = 1, ignore = ignore, print=F)

# empty vectors to be filled
sBPD_imp <- vector("list",5)
sBPD_imp_test <- vector("list",5)

# split imputed dataset into train and test based on the random vector
sBPD_train <- filter(sBPD_mice_out, !ignore)
sBPD_test <- filter(sBPD_mice_out, ignore)
```

```r
# store train sets and test set
for (i in 1:5){
  sBPD_imp[[i]] <- mice::complete(sBPD_train,i)
#  sBPD_imp_test[[i]] <- mice::complete(sBPD_test, i)
}

sBPD_imp_test <- mice::complete(sBPD_test, action = "long")

# correlation plot for variables
cor_mat <- cor(sBPD[,-c(1:3, 8:14, 16, 20, 22, 26, 28)], use = "complete.obs")
corrplot(cor_mat,tl.col = "black")

# missingness summary table
variable_names <- c("Fraction of Inspired Oxygen at 44 Weeks", "Peak Inspiratory Pressure at 44 Weeks",

missingness_sBPD <- sBPD %>%
  miss_var_summary() %>%
  filter(n_miss > 0) %>%
  mutate(variable_names = variable_names) %>%
  select(c(4,1:3))

colnames(missingness_sBPD) <- c("Variable Name", "Variable", "Number of Missing", "Percentage % of Missi

missingness_sBPD %>%
  select(-2) %>%
  kable(booktabs = TRUE, caption = "Missingness Summary") %>%
  kableExtra::kable_styling(font_size = 8, latex_options = c("repeat_header", "HOLD_position"))

# Lasso model with cross validation
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_or_Death~ .
                        + ventilation_support_level.36*med_ph.36
                        + ventilation_support_level_modified.44:p_delta.44
                        + ventilation_support_level_modified.44:med_ph.44
                        + inspired_oxygen.44:med_ph.44,
                        data = df[,-c(1,4:6,26)])[,-1]
  y.ord <- df$Trach_or_Death

  # Generate folds
  k <- 10
  set.seed(1) # consistent seeds
  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)
```

```r
  # Get coefficients
  coef <- coef(lasso_mod, lambda=lasso_mod$lambda.min)
  return (coef)
}

# averaging the coefficients on 5 imputed train set
lasso_coef1 <- lasso(sBPD_imp[[1]])
lasso_coef2 <- lasso(sBPD_imp[[2]])
lasso_coef3 <- lasso(sBPD_imp[[3]])
lasso_coef4 <- lasso(sBPD_imp[[4]])
lasso_coef5 <- lasso(sBPD_imp[[5]])
lasso_coef <- cbind(lasso_coef1, lasso_coef2, lasso_coef3,
                    lasso_coef4, lasso_coef5)
avg_coefs_lasso <- apply(lasso_coef, 1, mean)

# store the averaged set of coefficients and their occurrence of exclusion
variables_occurrence_lasso <-
  as.matrix((lasso_coef1 == 0) + (lasso_coef2 == 0) + (lasso_coef3 == 0) +
              (lasso_coef4 == 0) + (lasso_coef5 == 0)) %>%
  data.frame() %>%
  mutate(occurrence_zeros_lasso = s0, coefs_lasso = round(avg_coefs_lasso, 5)) %>%
  select(-s0)

# Best Subset with cross validation
bestsubset <- function(df) {
  #' Runs 10-fold CV for best subset 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_or_Death~ .
                        + ventilation_support_level.36*med_ph.36
                        + ventilation_support_level_modified.44:p_delta.44
                        + ventilation_support_level_modified.44:med_ph.44
                        + inspired_oxygen.44:med_ph.44,
                        data = df[,-c(1,4:6,26)])[,-1]
  y.ord <- df$Trach_or_Death

  # number of folds
  k <- 10

  # Best Subset model
  bs_mod_cv <- L0Learn.cvfit(x.ord, y.ord, nFolds = k, seed = 1,
                             penalty = "L0", loss = "Logistic", intercept = TRUE)
  bs_mod <- bs_mod_cv$fit

  # Get coefficients
  lambda.min <- which.min(bs_mod_cv$cvMeans[[1]]) # the index of lambda that yields minimum cv errors
  coef <- c(bs_mod_cv$fit$a0[[1]][lambda.min], bs_mod_cv$fit$beta[[1]][,lambda.min])
  return (coef)
}

# averaging the coefficients on 5 imputed train set
```

```r
bestsubset_coef1 <- bestsubset(sBPD_imp[[1]])
bestsubset_coef2 <- bestsubset(sBPD_imp[[2]])
bestsubset_coef3 <- bestsubset(sBPD_imp[[3]])
bestsubset_coef4 <- bestsubset(sBPD_imp[[4]])
bestsubset_coef5 <- bestsubset(sBPD_imp[[5]])
bestsubset_coef <- cbind(bestsubset_coef1, bestsubset_coef2, bestsubset_coef3,
                    bestsubset_coef4, bestsubset_coef5)
avg_coefs_bestsubset <- apply(bestsubset_coef, 1, mean)

# run this section to get all coefficient names as bestsubset does not return any names
# same code as in Lasso model, just with all interactions
x.ord <- model.matrix(Trach_or_Death~ (.)^2,
                          data = sBPD_imp[[1]][,-c(1,4:6,26)])[,-1]
y.ord <- sBPD_imp[[1]]$Trach_or_Death

# Generate folds
k <- 10
set.seed(1) # consistent seeds
folds <- sample(1:k, nrow(sBPD_imp[[1]]), 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, lambda=lasso_mod$lambda.min)
# store the averaged set of coefficients and their occurrence of exclusion
variables_occurrence_bs <-
  data.frame(occurrence_zeros_bs =
               as.matrix((bestsubset_coef1 == 0) + (bestsubset_coef2 == 0) + (bestsubset_coef3 == 0) +
                           (bestsubset_coef4 == 0) + (bestsubset_coef5 == 0)),
             coefs_bs = round(avg_coefs_bestsubset,5))

row.names(variables_occurrence_bs) <- names(avg_coefs_lasso)


variables_summary <- cbind(variables_occurrence_lasso, variables_occurrence_bs)

variables_summary[2:10,] %>%
  filter(!occurrence_zeros_lasso == 5 | !occurrence_zeros_bs == 5) %>%
  kable(booktabs = TRUE,
      caption = "Predictors at Birth level") %>%
  row_spec(c(2,4,8), background = "pink") %>%
  kableExtra::kable_styling(font_size = 8, latex_options = c("repeat_header", "HOLD_position", "scale_d

variables_summary[11:17,] %>%
  filter(!occurrence_zeros_lasso == 5 | !occurrence_zeros_bs == 5) %>%
  kable(booktabs = TRUE,
      caption = "Predictors at 36 wk PMA level") %>%
  row_spec(c(3,4), background = "yellow") %>%
  kableExtra::kable_styling(font_size = 8, latex_options = c("repeat_header", "HOLD_position", "scale_d
```

```r
variables_summary[18:24,] %>%
  filter(!occurrence_zeros_lasso == 5 | !occurrence_zeros_bs == 5) %>%
  kable(booktabs = TRUE,
      caption = "Predictors at 44 wk PMA level") %>%
  row_spec(c(1,3,4,6), background = "#00c19a") %>%
  kableExtra::kable_styling(font_size = 8, latex_options = c("repeat_header", "HOLD_position", "scale_d

variables_summary[25:31,] %>%
  filter(!occurrence_zeros_lasso == 5 | !occurrence_zeros_bs == 5) %>%
  kable(booktabs = TRUE,
      caption = "Interaction-Term Predictors") %>%
  row_spec(c(1,3,4), background = "#E68613") %>%
  kableExtra::kable_styling(font_size = 8, latex_options = c("repeat_header", "HOLD_position", "scale_d
# model matrix for test dataset
X.test <- model.matrix(Trach_or_Death~. , data = sBPD_imp_test[c(4:5, 9:27, 29)])

X.test2 <- model.matrix(Trach_or_Death~.
                        + ventilation_support_level.36*med_ph.36
                        + ventilation_support_level_modified.44:p_delta.44
                        + ventilation_support_level_modified.44:med_ph.44
                        + inspired_oxygen.44:med_ph.44, data = sBPD_imp_test[c(4:5, 9:27, 29)])


# predicted probability with three models
pred_lasso <- plogis(X.test %*% avg_coefs_lasso[-c(25:31)])
pred_lasso_interactions <- plogis(X.test2 %*% avg_coefs_lasso)
pred_bestsubset <- plogis(X.test %*% avg_coefs_bestsubset[-c(25:31)])
pred_bestsubset_interactions <- plogis(X.test2 %*% avg_coefs_bestsubset)

# roc for all three models
roc_lasso <- roc(sBPD_imp_test$Trach_or_Death, pred_lasso)
roc_lasso_interactions <- roc(sBPD_imp_test$Trach_or_Death, pred_lasso_interactions)
roc_bs <- roc(sBPD_imp_test$Trach_or_Death, pred_bestsubset)
roc_bs_interactions <- roc(sBPD_imp_test$Trach_or_Death, pred_bestsubset_interactions)


# evaluation metrics helper function
eval_metrics <- function(pred_prob, actual, threshold = 0.1){
  # predicted probabilities and actual values
  prediction <- ifelse(pred_prob > threshold, 1, 0)
  sensitivity <- sum(prediction == 1 & actual == 1) / sum(actual == 1)
  specificity <- sum(prediction == 0 & actual == 0) / sum(actual == 0)
  BS <- mean( (pred_prob - ifelse(actual == 1, 1, 0))^2 )
  AUC <- as.numeric(roc(sBPD_imp_test$Trach_or_Death, pred_prob)$auc)
  return(c(sensitivity, specificity, BS, AUC))
}

# evaluation metrics
metrics_lasso <- eval_metrics(pred_lasso, sBPD_imp_test$Trach_or_Death)
metrics_lasso_interactions <- eval_metrics(pred_lasso_interactions, sBPD_imp_test$Trach_or_Death)
metrics_bestsubset <- eval_metrics(pred_bestsubset, sBPD_imp_test$Trach_or_Death)
metrics_bestsubset_interactions <- eval_metrics(pred_bestsubset_interactions, sBPD_imp_test$Trach_or_Dea
```

```r
metrics_lasso
metrics_lasso_interactions
metrics_bestsubset
metrics_bestsubset_interactions
# put the evaluation metrics into a table
metrics_table <- data.frame(metrics_lasso, metrics_lasso_interactions, metrics_bestsubset, metrics_bests

rownames(metrics_table) <- c("Sensitivity", "Specificity", "Brier Score", "AUC")
colnames(metrics_table) <- c("Lasso", "Lasso with Interactions", "Best Subset", "Best Subset with Intera

metrics_table %>% kable(caption = "Model Evaluation Metrics", align = "c",booktabs = T) %>%
kable_styling(full_width=T,latex_options = c('HOLD_position'),font_size = 10)

# ROC Curves
plot(roc_lasso, main = "Multiple ROC Curves", col = "brown3", lwd = 2)
plot(roc_lasso_interactions, col = "brown1", add = TRUE, lwd=2)
plot(roc_bs, col = "navy", add = TRUE, lwd = 2)
plot(roc_bs_interactions, col = "cornflowerblue", add = TRUE, lwd=2)

legend("bottomright", legend = c("Lasso", "Lasso with Interactions", "Best Subset", "BS with Interactio

# calibration plot helper function
calibration_plot <- function(pred_prob) {
  num_cuts <- 10
  calib_data <- data.frame(prob = pred_prob,
                           bin = cut(pred_prob, breaks = num_cuts),
                           class = ifelse(sBPD_imp_test$Trach_or_Death == 1, 1, 0))

  calib_data <- calib_data %>%
    group_by(bin) %>%
    summarize(observed = sum(class)/n(),
              expected = sum(prob)/n(),
              se = sqrt(observed*(1-observed)/n()))

  calib_plot <-
    ggplot(calib_data) +
    geom_abline(intercept = 0, slope = 1, color="red") +
    geom_errorbar(aes(x = expected, ymin=observed-1.96*se,
                      ymax=observed+1.96*se),
                  colour="black", width=.01)+
    geom_point(aes(x = expected, y = observed)) +
    labs(x="Expected Proportion", y="Observed Proportion") +
    theme_minimal()

  return(calib_plot)
}

# calibration plots
ggarrange(calibration_plot(pred_lasso_interactions), calibration_plot(pred_bestsubset_interactions))

# try excluding `center` variable
pred_lasso.nocenter <- plogis(X.test[,-(2:9)] %*% avg_coefs_lasso[-(2:9)])
```

```r
dropcenter <- data.frame(eval_metrics(pred_lasso.nocenter, sBPD_imp_test$Trach_or_Death), metrics_table$
colnames(dropcenter) <- c("Lasso without Center Variable" , "Lasso with Center Variable")
rownames(dropcenter) <- rownames(metrics_table)

dropcenter %>% kable(caption = "Model Evaluation Metrics for Lasso with and without Center", align = "c
kable_styling(full_width=T,latex_options = c('HOLD_position'),font_size = 10)
```