Predictive Modeling for Tracheostomy and Mortality Outcomes in Bronchopulmonary Dysplasia

Abstract

This study focused on developing predictive models for tracheostomy and mortality outcomes in infants with Bronchopulmonary Dysplasia (BPD), leveraging data from the BPD Collaborative Registry. The registry comprised infants born before 32 weeks of gestational age, diagnosed with severe BPD. Data analysis involved exploring missing data patterns, where discrepancies in surfactant administration across various centers were noted, and implementing multiple imputation to handle missing-at-random data. The study identified key variables for outcomes through exploratory analysis, including high correlations and potential confounders like gestational age and positive end-expiratory pressure.

The dataset was split into training and testing sets, with a 70-30 proportion, to ensure robust model validation. Two types of models were developed for each outcome: one incorporating significant coefficients with a random effect from medical centers and another including interactions between time and clinical variables at 36 and 44 weeks. However, the addition of interaction terms did not significantly improve the models. The performance of the models was assessed using Area Under the Curve (AUC) and F1 scores, revealing satisfactory results for tracheostomy but poor F1 scores for death models, indicating an area for further model improvement.

The study's final multilevel model, chosen for its simplicity and effectiveness, combined significant variables and accounted for variations due to center-specific practices. While the models identified several key characteristics influencing tracheostomy and mortality, the study acknowledges limitations, including the absence of diverse interaction terms and a limited scope of random effects. Future research could address these limitations by exploring more interaction terms and broadening the scope of random effects. Despite these limitations, the study provides crucial insights into the factors influencing tracheostomy and mortality in neonates with BPD, guiding targeted interventions and policy decisions to improve neonatal care outcomes.

Keywords: Bronchopulmonary Dysplasia, Tracheostomy, Mortality, Predictive Modeling, Multilevel Models, Neonatal Care

1. Introduction

Bronchopulmonary Dysplasia (BPD) is a complication associated with prematurity, impacting a significant number of infants each year, particularly in its severe form, affecting 10,000-15,000 newborns annually. Various factors, including genetics and epigenetics, influence the development of BPD. This condition manifests as a persistent lung ailment, primarily afflicting prematurely born infants, necessitating oxygen therapy for their respiratory support. In BPD, there is notable damage to the lungs and airways (bronchi), leading to tissue damage (dysplasia) in the small air sacs of the lungs (alveoli). The severity of BPD is categorized into different grades, with Grade 3 BPD marking a critical point where reliance on a ventilator is necessary at 36 weeks corrected gestational age. Notably, 75% of infants with Grade 3 BPD continue to require ventilator support upon discharge, while 25% do not. For those who need ventilator support upon discharge, a tracheostomy involving a surgical opening in the neck facilitating connection to a ventilator becomes a prerequisite. The incidence of tracheostomy in infants with BPD ranges from 2-4%, escalating to 12% in cases of severe or Grade 3 BPD.

While the advantages of performing a tracheostomy include ensuring a stable airway, improving ventilator synchronization, and fostering growth, it is essential to acknowledge the associated risks. These risks include an increased likelihood of mortality compared to cases without tracheostomy, the potential for accidental decannulation leading to fatal outcomes, cannula obstruction with similar dire consequences, elevated rates of infection affecting the skin, trachea, and lungs, and the development of tracheal stenosis.

Given these considerations, cautious decision-making is crucial when contemplating the implementation of tracheostomy in infants diagnosed with BPD. Consequently, this study aims to develop statistical models utilizing clinical data collected at 36 and 44 weeks post-menstrual age (PMA). These models aim to predict the eventual necessity for tracheostomy or the likelihood of mortality preceding discharge, providing a valuable framework for informed decision-making in the management of BPD in newborns.

2. Data

Participants in this study were sourced from the BPD Collaborative Registry, a collaborative network of BPD programs in the United States and Sweden. The consortium was established to bridge evidence gaps and advance research for improving care in children affected by severe bronchopulmonary dysplasia (BPD). The registry focuses on infants born with a gestational age of less than 32 weeks and diagnosed with sBPD, defined according to the 2001 NHLBI criteria, specifically requiring FiO2 0.3 or positive pressure ventilation (invasive or non-invasive) at 36 weeks post-menstrual age (PMA). Standard demographic and clinical data are routinely

collected at four key time points: birth, 36 weeks PMA, 44 weeks PMA, and discharge. For this study, we extracted data from the registry for patients with BPD and complete growth information, covering the period from January 1 to July 19, 2021. At the time of analysis, 10 BPD Collaborative centers had contributed data that meets the study inclusion criteria

The dataset is structured around individual record_id entries representing premature infants. Key characteristics at birth include gender, corrected gestational age, originating center, birth measurements (weight, length, head circumference), and maternal characteristics such as race and ethnicity. Additional birth-related information, including delivery method, Prenatal Corticosteroids administration, Maternal Chorioamnionitis presence, and surfactant administration within the first 72 hours, is also captured.

Data at 36 and 44 weeks includes information on baby weight, Level of support, PEEP (Positive End-Expiratory and Pressure), Fraction of inspired O2, Peak inspiratory pressure, and medication administration for Pulmonary Hypertension. This dataset's primary outcomes of interest are whether infants underwent a tracheostomy at discharge and their mortality status.

3. Methodology

3.1. Analysis Plan

Given this study's dual outcomes of tracheostomy and death, we could make distinct models were constructed for each outcome. But in this study we will not create model for death outcome due to the small proportion of death cases in the dataset (only 5%). Concentrating on tracheostomy prediction was more urgent and feasible, as modeling rare events like death would require advanced techniques like resampling, which was beyond the scope of this study.

Initial steps involved an exploratory analysis of the dataset to identify visible patterns and aid in selecting significant variables pertinent to each outcome type. Furthermore, missing data was assessed, and suitable strategies for handling these gaps were determined.

Considering the many observed variables, a variable selection process was implemented to streamline the model, facilitating improved generalizability. The dataset was partitioned into two subsets – a training set and a testing set, with proportions of 0.7 and 0.3, respectively, to validate the model.

We will use two different types of dataset. One with only 36-week data and another including both 36-week and 44-week data, to determine if the 36-week data alone was sufficient for prediction or not. After that I will fit Lasso and Ridge regression model to both data. We regard the lasso and rigde model as the prediction model candidate and also as variables selection method. The selected variables from the best performing model will also be used to fit a multilevel model since we recognize that the dataset has multilevel structure. After that we will compare the performance of the Lasso, Ridge, and Multilevel model among the two dataset.

Given the binary nature of the outcomes, the performance of the models was assessed using the Area Under the Curve (AUC), F1 Score, Sensitivity, Specificity, and Precision. AUC is used to assess the general performance of the models in the dataset. Sensitivity is used to identifying true positives, crucial in avoiding missed tracheostomy needs. Precision is added to assess the ratio of true positives to predicted positives, an important factor considering the high cost of false positives leading to unnecessary procedures. Specificity is also included to ensure accurate identification of true negatives. F1 score is added to see the harmony of precision and sensitivity. This evaluation strategy ensures a comprehensive understanding of the model's accuracy and predictive capability.

3.2. Exploratory Data Analysis

The initial step involved checking for duplicate data, and we identified one record_id with duplicates, promptly removing them from the dataset. Moving on to missing data analysis, many missing values were observed, particularly in the dataset related to 44 weeks. The missing values for those variables reached approximately 40% for each, with the surfactant indicator also showing notable gaps. Upon closer inspection, it appeared that infants missing week 44 data were primarily those discharged before week 44. Discrepancies were also noted in surfactant administration, which was missing values across different centers. Centers 3, 5, and 12 had fewer missing values, while Centers 4 and 7 exhibited a higher likelihood. Divergent data recording practices were also apparent, with some centers showing low completion rates for 36 and 44-week data. Again, this reinforces our belief that the data is Missing-at-Random, and we think that multiple imputation is the appropriate method to impute the missing values.

Table 1: Missing Data Proportion for Each Variable

Variable	Observation Missing	Proportion Missing
inspired_oxygen.44	448	44.9799197
p_delta.44	448	44.9799197
weight_today.44	446	44.7791165
peep_cm_h2o.44	446	44.7791165
any_surf	433	43.4738956
ventilation_support_level.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.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

During univariate analysis, considering missing values, the dataset exhibited high imbalance, especially for death. Most infants did not undergo tracheostomy (85%) and were alive (95%). Most did not receive pulmonary hypertension medication at weeks 36 and 44 (93% and 83%). Prenatal corticosteroids were administered to most mothers (87%), and Maternal Chorioamnionitis was mostly absent (83%). Discharge after 44 weeks was predominant (64%), and most infants received surfactant in the first 72 hours (82%). Continuous data displayed outliers, such as unusually high birth weights (mean: 806, max: 2725), peak inspiratory pressures at 36 weeks (mean: 5.3, max: 46), peak inspiratory pressures at 44 weeks (mean: 7.6, max: 52), and discharge ages (mean: 52.8, max: 573.9).

Further examination identified variables with high correlation (above 75%), and one representative variable was retained to mitigate multicollinearity. We removed birth length and head circumference since they are highly correlated to birth weight. Additionally, variables with moderate correlation (around 35% - 75%) were noted for potential consideration, intending to incorporate ridge and lasso regularization during variable selection.

Summarizing covariates concerning outcomes (tracheostomy) revealed significant differences in some covariates between the two groups, suggesting potential confounders. Covariates with visually different boxplot distributions among outcome groups, such as gestational age and positive end-expiratory pressure (cm H2O) at 36 weeks, were considered potential predictors. Variables lacking significant differences, visual distinctiveness, and high correlation with other

variable were excluded from the initial variable selection. The variables dropped based on those reason are: maternal ethnicity, gestational age, birth length, birth head circumference, complete prenatal steroid, maternal chorioamnionitis, gender, surfactant, discharge time, death, and birth weight.

Below is the example of the summary table stratified by tracheostomy.

			Trach		
Variable	**N**	**Overall**, N = 996	**0**, N = 850	**1**, N = 146	**p-value**
center	986				
1		55 / 986 (5.6%)	32 / 844 (3.8%)	23 / 142 (16%)	
2		630 / 986 (64%)	566 / 844 (67%)	64 / 142 (45%)	
3		57 / 986 (5.8%)	56 / 844 (6.6%)	1 / 142 (0.7%)	
4		60 / 986 (6.1%)	49 / 844 (5.8%)	11 / 142 (7.7%)	
5		40 / 986 (4.1%)	35 / 844 (4.1%)	5 / 142 (3.5%)	
7		32 / 986 (3.2%)	31 / 844 (3.7%)	1 / 142 (0.7%)	
12		69 / 986 (7.0%)	34 / 844 (4.0%)	35 / 142 (25%)	
16		38 / 986 (3.9%)	37 / 844 (4.4%)	1 / 142 (0.7%)	
20		4 / 986 (0.4%)	4 / 844 (0.5%)	0 / 142 (0%)	
21		1 / 986 (0.1%)	0 / 844 (0%)	1 / 142 (0.7%)	
Unknown		10	6	4	
mat_race	940				0.010
0		538 / 940 (57%)	474 / 804 (59%)	64 / 136 (47%)	
1		290 / 940 (31%)	243 / 804 (30%)	47 / 136 (35%)	
2		112 / 940 (12%)	87 / 804 (11%)	25 / 136 (18%)	
Unknown		56	46	10	
mat_ethn	939				0.35
1		74 / 939 (7.9%)	66 / 803 (8.2%)	8 / 136 (5.9%)	
2		865 / 939 (92%)	737 / 803 (92%)	128 / 136 (94%)	
Unknown		57	47	10	
bw	996	806 (297)	814 (295)	761 (303)	0.004
ga	996	25.77 (2.14)	25.76 (2.14)	25.84 (2.13)	0.73
blength	918	32.5 (3.8)	32.6 (3.8)	32.0 (3.9)	0.15
Unknown		78	48	30	
birth_hc	919	23.19 (2.76)	23.22 (2.71)	22.99 (3.07)	0.12
Unknown		77	46	31	
del_method	993				0.035
1		285 / 993 (29%)	254 / 848 (30%)	31 / 145 (21%)	
2		708 / 993 (71%)	594 / 848 (70%)	114 / 145 (79%)	
Unknown		3	2	1	
prenat_ster	961				0.011
0		126 / 961 (13%)	118 / 830 (14%)	8 / 131 (6.1%)	
1		835 / 961 (87%)	712 / 830 (86%)	123 / 131 (94%)	
Unknown		35	20	15	
com_prenat_ster	803				0.97
0		193 / 803 (24%)	166 / 690 (24%)	27 / 113 (24%)	
1		610 / 803 (76%)	524 / 690 (76%)	86 / 113 (76%)	
Unknown		193	160	33	
mat_chorio	934				0.81
0		774 / 934 (83%)	662 / 800 (83%)	112 / 134 (84%)	
1		160 / 934 (17%)	138 / 800 (17%)	22 / 134 (16%)	
Unknown		62	50	12	
gender	992				>0.99
0		408 / 992 (41%)	348 / 846 (41%)	60 / 146 (41%)	
1		584 / 992 (59%)	498 / 846 (59%)	86 / 146 (59%)	

Unknown		4	4	0	
sga	981				0.005
0		778 / 981 (79%)	678 / 839 (81%)	100 / 142 (70%)	
1		203 / 981 (21%)	161 / 839 (19%)	42 / 142 (30%)	
Unknown		15	11	4	
any_surf	563				0.14
0		102 / 563 (18%)	93 / 488 (19%)	9 / 75 (12%)	
1		461 / 563 (82%)	395 / 488 (81%)	66 / 75 (88%)	
Unknown		433	362	71	
weight_today.36	904	2,121 (413)	2,132 (410)	2,024 (432)	0.025
Unknown		92	38	54	
ventilation_support_level.36	966				<0.001
0		117 / 966 (12%)	111 / 839 (13%)	6 / 127 (4.7%)	
1		589 / 966 (61%)	560 / 839 (67%)	29 / 127 (23%)	
2		260 / 966 (27%)	168 / 839 (20%)	92 / 127 (72%)	
Unknown		30	11	19	
inspired_oxygen.36	904	0.34 (0.15)	0.32 (0.13)	0.49 (0.20)	<0.001
Unknown		92	37	55	
p_delta.36	868	5 (10)	4 (9)	15 (12)	<0.001
Unknown		128	63	65	
peep_cm_h2o.36	879	6.3 (2.9)	6.2 (2.9)	7.5 (2.8)	<0.001
Unknown		117	59	58	
med_ph.36	966				<0.001
0		900 / 966 (93%)	798 / 839 (95%)	102 / 127 (80%)	
1		66 / 966 (6.8%)	41 / 839 (4.9%)	25 / 127 (20%)	
Unknown		30	11	19	
weight_today.44	550	3,646 (682)	3,667 (665)	3,550 (750)	0.23
Unknown		446	399	47	
ventilation_support_level.44	572				<0.001
0		269 / 572 (47%)	262 / 461 (57%)	7 / 111 (6.3%)	
1		146 / 572 (26%)	128 / 461 (28%)	18 / 111 (16%)	
2		157 / 572 (27%)	71 / 461 (15%)	86 / 111 (77%)	
Unknown		424	389	35	
inspired_oxygen.44	548	0.34 (0.15)	0.32 (0.13)	0.44 (0.19)	<0.001
Unknown		448	398	50	
p_delta.44	548	8 (14)	5 (12)	21 (16)	<0.001
Unknown		448	397	51	
peep_cm_h2o.44	550	4.3 (4.5)	3.4 (4.1)	8.7 (3.6)	<0.001
Unknown		446	396	50	
med_ph.44	572				<0.001
0		473 / 572 (83%)	413 / 461 (90%)	60 / 111 (54%)	
1		99 / 572 (17%)	48 / 461 (10%)	51 / 111 (46%)	
Unknown		424	389	35	
hosp_dc_ga	872	53 (27)	49 (24)	80 (30)	<0.001
Unknown		124	86	38	
death	994				<0.001
0		940 / 994 (95%)	811 / 848 (96%)	129 / 146 (88%)	
1		54 / 994 (5.4%)	37 / 848 (4.4%)	17 / 146 (12%)	
Unknown		2	2	0	

[1] n / N (%); Mean (SD)

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

3.3. Data splitting

We split the data before applying any methods we planned to use. This is done to avoid data

leaking issues that would lead to overfitting. The data training and data testing proportions that will be used in this setting are 0.7 and 0.3. Also, we tried to separate the data that only contained 36-weeks data and the data that contained 36 and 44 weeks data because we seek to determine if the 36-week data alone was sufficient for prediction and also because we do not want to impute missing values in the 44 weeks from the data that does not have any 44 weeks of data. These data will undergo missing data imputation and variable selection separately.

3.4. Missing Data Impulation

In our dataset, missing values were predominantly assumed to follow a Missing-at-Random (MAR) pattern. Addressing this, we employed multiple imputation techniques, specifically using the 'mice' function from the MICE package. Our parameter settings for this function were m = 5 to generate five distinct imputed datasets, with other parameters left at their default settings. The imputation was separately conducted for each outcome-driven dataset (tracheostomy and death), further divided into subsets of 36-week data and a combination of 36 and 44-week data, each inclusive of base characteristics.

3.5. Variable and Model Selection Process

The process of selecting variables and models was methodical and based on preliminary findings from exploratory data analysis. The initial selection was guided by identifying variables that exhibited significant distributional differences or noticeable visual disparities across different outcome groups. We employed a dual approach using Lasso and Ridge regression techniques to refine the model selection. Lasso regression was chosen for its capacity to reduce the influence of less significant variables to zero, thereby simplifying the model. Ridge regression was utilized to address potential multicollinearity issues, a concern due to the suspected correlations between variables. This combination aimed to balance model simplicity and accuracy, ensuring both interpretability and generalizability of the model.

The optimal lambda values for both Lasso and Ridge regressions was determined through a cross-validation process. This process incorporated the five imputed datasets generated during our multiple imputation phase. By fitting a Lasso/Ridge regression model to each dataset, we accounted for the uncertainty inherent in the missing data, reducing bias in our final model. Cross-validation played a crucial role in this phase, helping to fine-tune the lambda parameter in both regression models, optimizing for minimal prediction error and exclusion of non-essential effects.

We calculated the AUC, F1, Sensitivity, Specificity, and Precision for both the Lasso and Ridge models using the test dataset. These results are presented in Figure 1 below. For tracheostomy outcomes, the Lasso model demonstrated superior performance in all the metrics compared to Ridge model in the dataset that include 44 week. Ridge perform better in some metrics like Sensitivity and F1 in the dataset that only has baseline and 36 week data. But overall, we see that Lasso in the 44 week dataset has the best perfromance. It seemed that only using

36 weeks data is not sufficient for making good prediction model. From the best lasso model, we found that there are two variables with 0 beta coefficient: small baby indicator and the delivery method. We excluded these variables in the multilevel model that we built after this process.

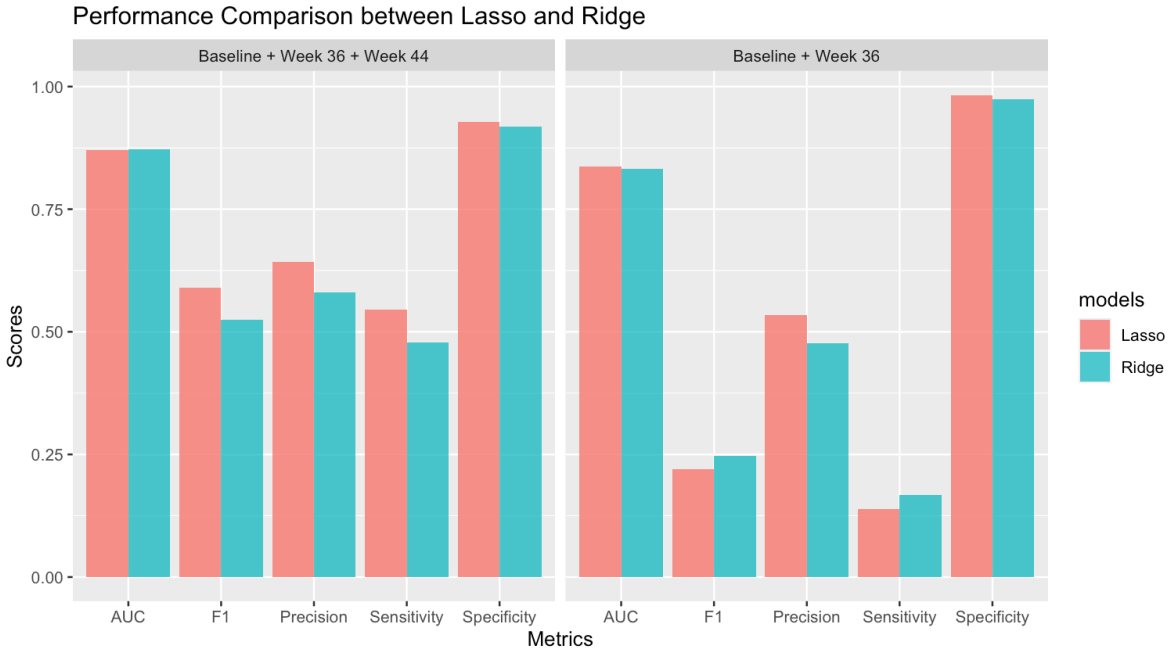

Figure 1: Metrics Comparison between Lasso and Ridge

In the development of our multilevel models, we incorporated significant variables identified from our previous lasso model analysis and introduced a random effect associated with medical centers. In addition, we incorporated time as the fixed effect. This model aimed to capture variations attributable to center-specific practices and characteristics.

Upon validation using the test dataset, we compare the perfromance of the multilevel model with the lasso and rigde. Although all three models showed good results in AUC and specificity (expected, because we have a lot of non event outcome), lasso performed best in sensitivity and precision – key for avoiding unnecessary or missed tracheostomies. Therefore, we chose the lasso model as the final model for this study. The best lasso model had 0.87 AUC, 0.59 F1 Score, 0.55 Sensitivity, 0.93 Specificity, and 0.64 Precision. We noted that our best model has mediocre sensitivity, precision, and F1 score. We suspected that it is due to the high imbalance in the outcome proportion in which we only had around 15% tracheostomy event happened.

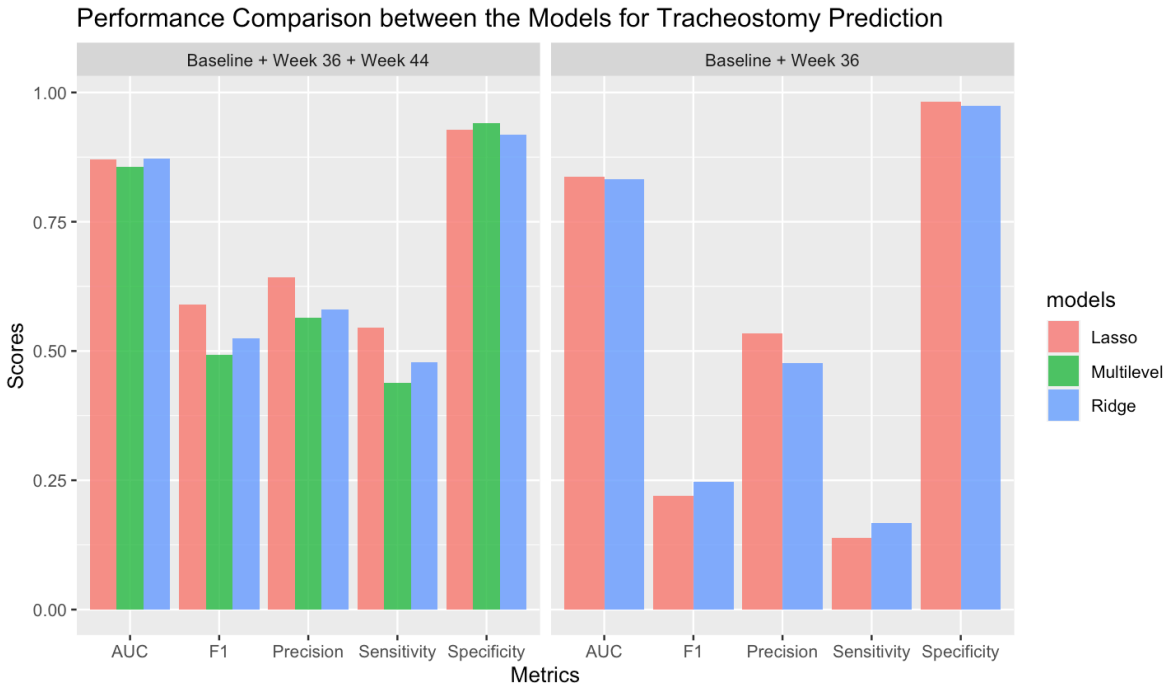

Figure 2: Metrics Comparison between Lasso and Ridge

4. Final model Result and Discussion

The lasso model was chosen as our final model. To finalize our analysis, we repeated the imputation process and the construction of the lasso model, this time employing the full dataset. The coefficients from our final model are listed and discussed in detail below.

Table 2: Estimate of Lasso Coefficients

	Estimate
prenat_ster1	2.175
ventilation_support_level.362	1.993
ventilation_support_level.diff	1.319
med_ph.361	1.259
mat_race2	1.237
inspired_oxygen.36	1.186
med_ph.diff	0.573
mat_race1	0.558
inspired_oxygen.diff	0.316
peep_cm_h2o.36	0.143
peep_cm_h2o.diff	0.067
p_delta.diff	0.000
p_delta.36	-0.014
ventilation_support_level.361	-0.016
weight_today.36	-0.053
weight_today.diff	-0.251
(Intercept)	-5.493

In the case of tracheostomy, our model identified several key characteristics associated with a higher odds of this intervention.

In the case of tracheostomy, our model identified several key characteristics associated with a higher likelihood of this intervention. Notably, Baby whose mother had Prenatal Corticosteroids has significantly higher odds of having tracheostomy compared to baby whose mother did not have Prenatal Corticosteroids. It is the same case with baby that has Invasive positive pressure support level on week 36, they have higher odds compared to baby who has non respiratory support at week 36. These findings suggest a complex interplay of institutional, prenatal, and clinical care factors in the likelihood of tracheostomy among neonates.

In conclusion, our models reveal that tracheostomy in neonates are influenced by a combination of hospital-specific factors, prenatal conditions, and specific medical interventions. The identification of these factors is crucial for understanding the clinical pathways and can guide targeted interventions and policy decisions to improve neonatal care outcomes.

5. Limitation

This study, while providing valuable insights, has several limitations that should be acknowledged.

1. Little to none Interaction Terms

Another limitation is the almost-absence of interaction terms in our model. We only tried to get the interaction terms between the time and the variables at 36 and 44 weeks but it could be that there is exist better interaction terms. Interaction terms can often reveal complex

interdependencies between variables that are not discernible when variables are considered independently. The inclusion of such interactions could potentially enhance the predictive accuracy and explanatory power of the model.

2. Limited Scope of Random Effects

Finally, the model's scope of random effects was restricted to only the medical center. This limited approach may overlook other significant random effects that could influence outcomes, such as patient demographics, staff characteristics, or temporal factors. Broadening the scope of random effects to include these additional variables could capture a more comprehensive range of influences, providing a more accurate and generalizable model.

In conclusion, while our study offers important findings, the aforementioned limitations suggest avenues for future research to enhance the model's comprehensiveness and accuracy.

Supplemental Material

Supplemental material can be seen in this github page

Code Appendix

```
##############
### SETUP ###
##############

library(formatR)

knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(message = F)
knitr::opts_chunk$set(warning = F)
knitr::opts_chunk$set(fig.align="center")
knitr::opts_chunk$set(fig.width=8, fig.height=6)

###############
### LIBRARY ###
###############

library(tidyverse)
library(ggplot2)
library(naniar)
library(gt)
library(gtsummary)
library(kableExtra)
library(GGally)
library(corrplot)
library(patchwork)
library(splitstackshape)
library(mice)
library(glmnet)
library(pROC)
library(lme4)
#####################################
### READ DATA AND PREPROCESSING ###
#####################################

# Set working directory and read data
setwd("/Users/amirahff/Documents/Brown Biostatistics/PHP 2550/project2")
raw_df <- read.csv("project2.csv")

# Check data type
```

```r
str(raw_df)

# Is there any duplicate data
# Yes, record_id = 2000824
raw_df %>%
  group_by(record_id) %>%
  count() %>%
  filter(n > 1)

# Remove duplicate data
df = raw_df %>%
  group_by(record_id) %>%
  distinct() %>%
  ungroup()

# Rename some columns and change data type
df = df %>%
  rename('peep_cm_h2o.36' = 'peep_cm_h2o_modified.36'
         ,'peep_cm_h2o.44' = 'peep_cm_h2o_modified.44'
         ,'ventilation_support_level.44' = 'ventilation_support_level_modified.44'
         ,'trach' = 'Trach'
         ,'death' = 'Death') %>%
  mutate(prenat_ster = case_when(prenat_ster=='Yes'~1,prenat_ster=='No'~0)
         ,com_prenat_ster = case_when(com_prenat_ster=='Yes'~1,com_prenat_ster=='No'~0)
         ,mat_chorio = case_when(mat_chorio=='Yes'~1,mat_chorio=='No'~0)
         ,gender = case_when(gender=='Male'~1,gender=='Female'~0)
         ,sga = case_when(sga=='SGA'~1,sga=='Not SGA'~0)
         ,any_surf = case_when(any_surf=='Yes'~1,any_surf=='No'~0)
         ,death = case_when(death=='Yes'~1,death=='No'~0)) %>%
  mutate(record_id = as.factor(record_id)
         , center = as.factor(center)
         , mat_race = as.factor(mat_race)
         , mat_ethn = as.factor(mat_ethn)
         , del_method = as.factor(del_method)
         , prenat_ster = as.factor(prenat_ster)
         , com_prenat_ster = as.factor(com_prenat_ster)
         , mat_chorio = as.factor(mat_chorio)
         , gender = as.factor(gender)
         , sga = as.factor(sga)
         , ventilation_support_level.36 = as.factor(ventilation_support_level.36)
         , med_ph.36 = as.factor(med_ph.36)
```

```
         , ventilation_support_level.44 = as.factor(ventilation_support_level.44)
         , med_ph.44 = as.factor(med_ph.44)
         , any_surf = as.factor(any_surf)
         , trach = as.factor(trach)
         , death = as.factor(death)) %>%
  mutate(across(where(is.numeric), round, 4))

####################
### MISSING DATA ###
####################

# Variables' missing data proportion
varMissingProp = miss_var_summary(df)
varMissingProp %>%
  filter(n_miss > 0) %>%
  kableExtra::kbl(caption = 'Missing Data Proportion for Each Variable'
                  , booktabs = T
                  , escape = T
                  , align = 'c'
                  , col.names = c('Variable','Observation Missing','Proportion
                                  Missing')) %>%
  kableExtra::kable_classic(full_width = F
                            , html_font = 'Cambria'
                            , font_size = 6
                            , latex_options = 'HOLD_position')
############################
### COMPARE TRACH GROUP ###
############################

#Summary by Trach
df %>%
  dplyr::select(-c(record_id)) %>%
  tbl_summary(by = trach
              , statistic = list(
                all_continuous() ~ "{mean} ({sd})"
                ,all_categorical() ~ "{n} / {N} ({p}%)"
                )) %>%
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 2)) %>%
  add_overall() %>%
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
```

```
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Trach**") %>%
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE, longtable = TRUE) %>%
  kableExtra::kable_classic(full_width = F
                            , html_font = 'Cambria'
                            , font_size = 7
                            , latex_options = 'scale_down')
load('final_coef.Rda')
round(final_coef,3) %>%
  arrange(desc(Estimate)) %>%
  kableExtra::kbl(caption = 'Estimate of Lasso Coefficients'
                  , booktabs = T
                  , escape = T
                  , align = 'c') %>%
  kableExtra::kable_classic(full_width = F
                            , font_size = 7
                            , html_font = 'Cambria'
                            , latex_options = 'HOLD_position')
```