Project 2: Predicting Tracheostomy/Death in Neonates with Severe BPD Using Lasso Logistic Regression

Mavis (Xinwen) Liang

1 Abstract

Accurate prediction of the need for tracheostomy at early postmenstrual ages (PMA) can have significant implications for families and clinical decision-making. In this analysis, we utilized a national dataset containing demographic, diagnostic, and respiratory parameters of infants with sBPD to develop a logistic regression model to predict the composite outcome of tracheostomy or death in neonates with severe bronchopulmonary dysplasia (sBPD), with a focus on assessing the impact of missing data and accounting for patients discharged before or after 44 week. The model has an overall good performance when we validate it using the test se, and has the potential implications for clinical practice in predicting Tracheostomy/Death.

The code for this analysis can be found in Github: https://github.com/Mavis-Liang/pda_project2.

2 Introduction

This collaborative project, undertaken in partnership with Dr. Chris Schmid from the Biostatistics Department at Brown University, aims to address a critical gap in understanding the indication criteria and timing of tracheostomy placement in neonates with severe bronchopulmonary dysplasia (sBPD). The project is motivated by the lack of clarity surrounding the optimal timing for tracheostomy placement in this population and the potential benefits of early intervention for growth.

Data for this study were sourced from the BPD Collaborative Registry, a multi-center consortium comprising interdisciplinary BPD programs in the United States and Sweden. The registry includes infants with a gestational age of less than 32 weeks and diagnosed with severe bronchopulmonary dysplasia (sBPD) based on 2001 NHLBI criteria. The study specifically

considers infants requiring FiO2 0.3 or positive pressure ventilation (invasive or non-invasive) at 36 weeks PMA. Standard demographic and clinical data are collected at four key time points: birth, 36 weeks PMA, 44 weeks PMA, and discharge.

The inclusion criteria for this study involve patients with BPD and complete growth data, and the data query was conducted for patients meeting these criteria between January 1 and July 19, 2021. The analysis incorporates data contributed by 10 BPD Collaborative centers, reflecting a diverse and representative cohort.

Throughout the statistical analysis report, we built a lasso logistic regression model that can account for patients discharge early or later altogether, after missing value imputations. This model has the potential implications for clinical practice and patient outcomes.

3 Methods

3.1 Data Preprocessing

Variable selection and transformation

The two outcomes (Tracheotomy and Death) are combined into one singel outcome of `trach_or_death`, since the proportion of death is too low in the data, and, on the other hand, both tracheotomy and death can indicate a severe outcome of sBPD in newborns.

The covariates and how they were processed are summarized as below:

Categorical variables: (1) Medical center: the medical center of number 21 has only one data point, thus we dropped this data point. (2) Maternal race. (3) Maternal Ethnicity. (4) Delivery Method. (5) Prenatal Corticosteroids. (6) Complete Prenatal Steroids: it's highly incomplete and closely associated with Prenatal Corticosteroids, thus we dropped this variable in our model. (7) Gender. (8) Was the infant small for gestational age. (9) Did the infant receive surfactant at any point in the first 72 hours? : We dropped this cavariates due to its high incompleteness. (10) Ventilation support level at 36 weeks. (11) Medication for Pulmonary Hypertension at 36 weeks. (12) Ventilation support level at 44 weeks. (13) Medication for Pulmonary Hypertension at 44 weeks.

Contiuous variables: (1) Birth weight (g): transformed into linear spline terms at the knot of 1610. (2) Obstetrical gestational age. (3) Birth length (cm): transformed into linear spline terms at the knot of 40. (4) Birth head circumference (cm). (5) Weight at 36 weeks: : transformed into linear spline terms at the knots of 1510 and 2710. (6) Fraction of Inspired Oxygen at 36 weeks: : transformed into linear spline terms at the knot of 0.789. (7) Peak Inspiratory Pressure (cmH2O) at 36 weeks. (8) Positive and exploratory pressure (cm H2O) at 36 weeks. (9) Weight at 44 weeks: transformed into linear spline terms at the knot of 2110. (10) Fraction of Inspired Oxygen needed at 44 weeks: transformed into linear spline terms at

the knot of 0.737. (11) Peak Inspiratory Pressure (cmH2O) needed at 44 weeks (12) Positive end exploratory pressure (cm H2O) needed at 44 weeks.

Multiple imputation

To deal with the missing patterns using imputation algorithms, we first divided the population into two cohorts, `Cohort 36` and `Cohort 44`, based on if their hospital discharge age is before or after 44 week. Those with missing values in hospital discharge age are categorized into `Cohort 36` or `Cohort 44` based on if they have at least one valid 44 week's measurement. Then, separately for the two cohorts, we employed Multiple Imputation by Chained Equations (MICE) with 5 imputations to handle missing data in both cohorts. We then combined the two cohorts, and formed 5 imputation datasets, for all individuals. In this way, instead of imputaing altogether, we don't have to impute the 44 week's measurements for `Cohort 36`, and we are only using the `Cohort 44` to impute those who should have a record at the 44th week.

3.2 Model Building

Separately for each of the 5 imputation dataset but with the same seed, we randomly sample 70% of the individuals to form a training set. We first fitted a multivariate logistic regression in one of the training set to find the preliminary significant levels of the variables.

To enhance the robustness of our coefficient estimates and account for missing data, on each training set, we fitted LASSO logistic regressions and employed 10-fold cross-validation to determine the optimal penalty factor for each LASSO model. The penalty factor was selected to minimize the Mean Square Error across the 10 folds. Then we pooled the coefficients obtained from the 5 best models corresponding to the respective imputed training sets by averaging coefficients across the 5 best models. The pooled coefficients work as our final model.

In the Lasso logistic regressions, we modeled the log odds of experiencing tracheotomy or death as a linear combination of the specified covariates from the Data Processing session. Instead of directly adding the 44 week's measurements themselves, we created an indicator variable of "whether a patient has been discharged at 44 week" and include this and its interaction term with the measurements at the 44th week in the model. In this way, the coefficients we get for most of the covariates are estimated with the whole data, while coefficients pertaining to measurements at the 44th week are specifically derived from patients who have not been discharged at that time.

3.3 Model evaluation

The 30% left-out samples from the 5 imputation sets were combined into a long dataset to form the test set (each individuals have 5 entries). We then used the pooled coefficients

and the test set to evaluate the model performance. We constructed the Receiver Operating Characteristic (ROC) curve depicting the trade-off between sensitivity and specificity across different probability thresholds, along with the AUC which quantify the model's discriminatory power. We also assess the agreement between the predicted probabilities against observed outcomes using calibration plot.

4 Results

4.1 Characteristics of the data

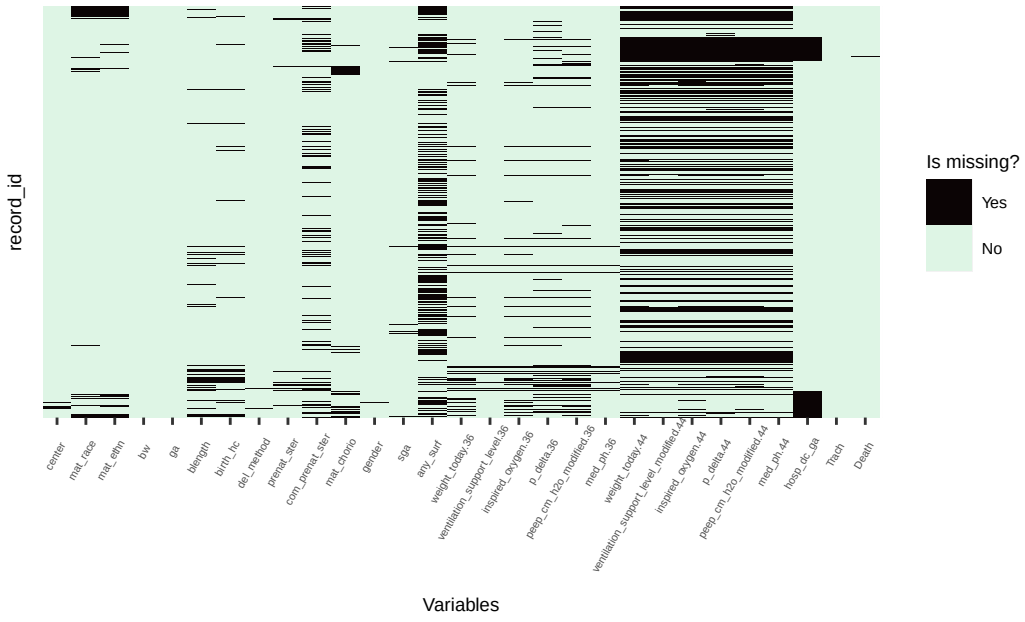

Figure 1: Missing patterns in original dataset

The data contains 993 individuals of 30 variables (after duplicated entries is removed for the individual with the recod id of 2000824, and center 21 is removed). For each individual, we have mother's demographic record, the infants' physical measurements, medical treatments and measurements at 36 week and at 44 week, and the outcome of Tracheotomy or Death. Figure 1 visualized the individual missingness. The data contains a high degree of missingness, which needs imputation, but will still have a large noise in our model and prediction. Specifically, measurements at the 44th week are highly incomplete due to the early discharge of the patients. Within patients discharged after 44 week, we still have ~90 over 685 missingness. Other

variables of Complete Prenatal Steroids and Did the infant receive surfactant at any point in the first 72 hours? also have large proportions of missing values.

Table 1: Summary of variables by outcome

	Tracheostomy or Death			
Variables	**Overall**, N = 993	**No**, N = 811	**Yes**, N = 182	**p-value**
Medical Center				
1	55 (100%)	25 (45%)	30 (55%)	
2	629 (100%)	545 (87%)	84 (13%)	
3	57 (100%)	55 (96%)	2 (3.5%)	
4	59 (100%)	47 (80%)	12 (20%)	
5	40 (100%)	33 (83%)	7 (18%)	
7	32 (100%)	31 (97%)	1 (3.1%)	
12	69 (100%)	28 (41%)	41 (59%)	
16	38 (100%)	37 (97%)	1 (2.6%)	
20	4 (100%)	4 (100%)	0 (0%)	
missing	10	6	4	
Maternal Race				0.002
0	536 (100%)	457 (85%)	79 (15%)	
1	290 (100%)	232 (80%)	58 (20%)	
2	111 (100%)	80 (72%)	31 (28%)	
missing	56	42	14	
Maternal Ethnicity				0.2
Hispanic or Latino	73 (100%)	64 (88%)	9 (12%)	
Not Hispanic or Latino	863 (100%)	703 (81%)	160 (19%)	
missing	57	44	13	
Birth Weight (g)	806 (297)	817 (285)	757 (341)	<0.001
Obstetrical Gestational Age	26 (2)	26 (2)	26 (2)	0.5
Birth Length (cm)	32 (4)	33 (4)	32 (4)	0.002
missing	78	45	33	
Birth Head Circumference (cm)	23.19 (2.76)	23.25 (2.65)	22.89 (3.29)	0.012
missing	77	44	33	
Delivery Method				0.019
Cesarean section	706 (100%)	564 (80%)	142 (20%)	
Vaginal delivery	284 (100%)	245 (86%)	39 (14%)	
missing	3	2	1	
Prenatal Corticosteroids				0.026
No	126 (100%)	113 (90%)	13 (10%)	
Yes	832 (100%)	679 (82%)	153 (18%)	
missing	35	19	16	
Complete Prenatal Steroids				0.8
No	192 (100%)	159 (83%)	33 (17%)	
Yes	608 (100%)	499 (82%)	109 (18%)	
missing	193	153	40	
Maternal Chorioamnionitis				>0.9
No	771 (100%)	633 (82%)	138 (18%)	
Yes	160 (100%)	132 (83%)	28 (18%)	
missing	62	46	16	
Gender				0.7
Female	406 (100%)	334 (82%)	72 (18%)	
Male	583 (100%)	473 (81%)	110 (19%)	
missing	4	4	0	
Small for Gestational Age				<0.001
Not SGA	775 (100%)	658 (85%)	117 (15%)	
SGA	203 (100%)	142 (70%)	61 (30%)	
missing	15	11	4	
Surfactant Received				0.095
No	101 (100%)	89 (88%)	12 (12%)	
Yes	461 (100%)	374 (81%)	87 (19%)	
missing	431	348	83	
Weight at 36 Weeks	2,121 (413)	2,142 (393)	1,986 (505)	<0.001
missing	92	32	60	
ventilation_support_level.36				<0.001
Invasive positive pressure	259 (100%)	140 (54%)	119 (46%)	
No respiratory support or supplemental oxygen	116 (100%)	109 (94%)	7 (6.0%)	
Non-invasive positive pressure	588 (100%)	553 (94%)	35 (6.0%)	
missing	30	9	21	
Inspired Oxygen at 36 Weeks	0.34 (0.15)	0.31 (0.12)	0.49 (0.21)	<0.001
missing	91	32	59	
Peak Inspiratory Pressure at 36 Weeks	5 (10)	4 (8)	16 (12)	<0.001
missing	128	50	78	
PEEP* at 36 Weeks	6 (3)	6 (3)	7 (3)	<0.001
missing	117	48	69	
Medication for PH* at 36 Weeks				<0.001
No	899 (100%)	770 (86%)	129 (14%)	
Yes	64 (100%)	32 (50%)	32 (50%)	

missing	30	9	21	
Weight at 44 Weeks	3,648 (681)	3,695 (643)	3,480 (781)	0.012
missing	444	383	61	
Ventilation Support at 44 Weeks				<0.001
Invasive positive pressure	157 (100%)	53 (34%)	104 (66%)	
No respiratory support or supplemental oxygen	269 (100%)	261 (97%)	8 (3.0%)	
Non-invasive positive pressure	145 (100%)	124 (86%)	21 (14%)	
missing	422	373	49	
Inspired Oxygen at 44 Weeks	0.34 (0.15)	0.31 (0.11)	0.46 (0.20)	<0.001
missing	446	382	64	
Peak Pressure at 44 Weeks	8 (14)	4 (11)	22 (16)	<0.001
missing	446	379	67	
PEEP at 44 Weeks	4 (4)	3 (4)	9 (3)	<0.001
missing	444	378	66	
Medication for PH at 44 Weeks				<0.001
No	473 (100%)	405 (86%)	68 (14%)	
Yes	98 (100%)	33 (34%)	65 (66%)	
missing	422	373	49	
Hospital Discharge Gestational Age	53 (27)	49 (24)	73 (30)	<0.001
missing	123	78	45	
Tracheostomy				<0.001
No	848 (100%)	811 (96%)	37 (4.4%)	
Yes	145 (100%)	0 (0%)	145 (100%)	
Death				<0.001
No	939 (100%)	811 (86%)	128 (14%)	
Yes	54 (100%)	0 (0%)	54 (100%)	

[1] PEEP: Positive end exploratory pressure; PH: Pulmonary Hypertension
[2] Pearson's Chi-squared test; Wilcoxon rank sum test

Table 1 provides a comprehensive summary of the variables stratified by the outcome of Tracheostomy or Death in a cohort of 993 individuals, along with the Pearson's Chi-squared test or Wilcoxon rank sum test p-values within each variables. Specifically, Medical Center 2 has a large proportion of patitient in our study, therefore, the effect of Medical Center 2 will have a high effect in model building. This missingness in the variables of Complete Prenatal Steroids and Surfactant Received is notable. Patitiens only have the outcome of either Tracheostomy or Death.

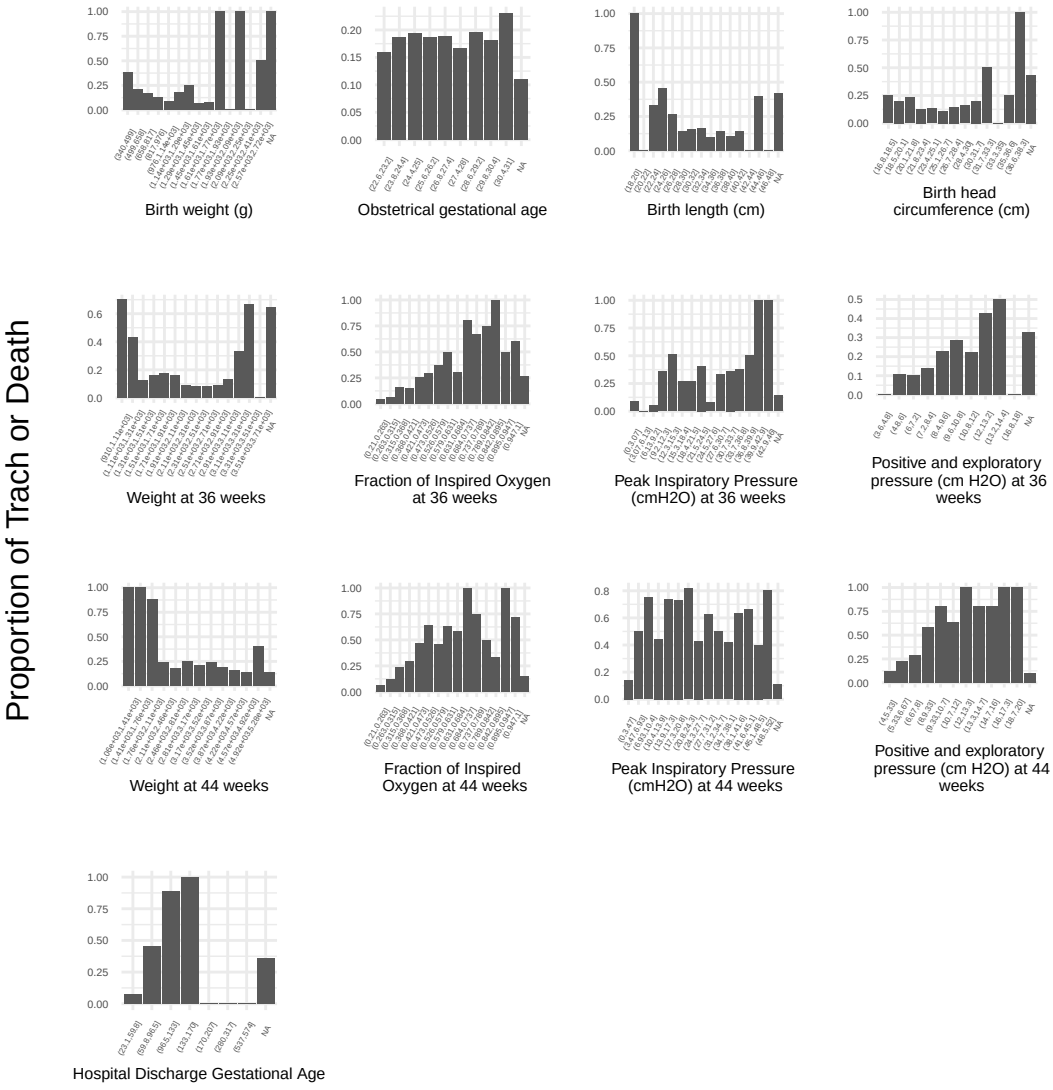

Figure 2: Proportion of Tracheotomy/Death in Continuous Variables

Figure 2 illustrates how the proportion of Tracheotomy/Death has been changing in response to variations in continuous variables. The continuous variables are binned, and within each bin, the proportion of Tracheotomy/Death is computed. As we see, the observed relationships are not linear. Rather, the propotions along the x-axis are often in a concave shape, i.e. high proportions of Tracheotomy/Death often occur at extreme values. Specifically, for the variables measuring the Fraction of Inspired Oxygen at 36 weeks and Fraction of Inspired Oxygen at 44 weeks, the proportion of Tracheotomy/Death first ascends then declined, indicating a linear

spline patterns of the outcomes to the variables.

4.2 Model Interpretation

Table 2 shows the pooled coeficients estimated with cross-validation lasso logistic regression in 5 imputated training set. The lasso coefficients being positive indicates a higher odds of getting Tracheotomy/Death comparing to the reference group or as the continous variable grow by one unit. In this sense, patients in center 12 with maternal race 2, having prenatal steroids, and being small, having a Ventilation support level 2 at 36 or 44, weeks, and having any Medication for Pulmonary Hypertension at 44 weeks, would have a higher odds of getting Tracheotomy/Death. For patients with a birth weight higher than 1610g, the odds of getting Tracheotomy/Death will increase by 12.835 times as the birth weight increased by 1 unit. Same interpretations can be drawn for the coefficients that have a negative value.

From the last 10 variables which include the indicator of whether the patient has been discharged at the week of 44 and its interaction with other 44 week's measurements, we see that the measurements at the 44 week are important predictors for predicting Tracheotomy/Death. Specifically, the variable of Ventilation support level 2 at 44 weeks and Medication for Pulmonary Hypertension at 44 weeks have coefficients of 1.807 and 1.006, which can change our predictions in a relatively large scale.

On the other hand, variables of Obstetrical gestational age, Birth head circumference (cm), Gender, Peak Inspiratory Pressure (cmH2O) at 36 weeks, Positive and exploratory pressure (cm H2O) at 36 weeks, Peak Inspiratory Pressure (cmH2O) needed at 44 weeks and Weight at 44 weeks are not useful in predicting Tracheotomy/Death, since their coefficients shrinked to very close to 0s in LASSO regression.

To use this model for prediction, we first identify whether a patient is discharged before 44 week. For patients discharged before 44 week, we only use the coefficients and variables above `disc_44` to estimate the probibility. For patients discharged after 44 week, we utilize their measurements at the 44 week. Thus for these patients we use all the variables in Table 2 for estimation.

Table 2: Association of selected variables with tracheostomy and death

	Lasso coefficient	Odds ratio
Intercept	-3.825	0.022
center2	-0.741	0.477
center3	-0.571	0.565
center4	-0.516	0.597
center5	0.052	1.054
center7	-0.512	0.600
center12	1.091	2.977
center16	-0.148	0.862
center20	-0.970	0.379
mat_race1	0.048	1.049
mat_race2	0.044	1.045
mat_ethn2	0.191	1.211
ga	0.004	1.004
birth_hc	0.000	1.000
del_method2	0.368	1.444
prenat_sterYes	0.600	1.821

mat_chorioYes	-0.024	0.976
genderMale	0.000	1.000
sgaSGA	0.172	1.187
ventilation_support_level.361	-0.198	0.820
ventilation_support_level.362	0.748	2.113
p_delta.36	0.001	1.002
peep_cm_h2o_modified.36	0.000	1.000
med_ph.361	0.034	1.035
hosp_dc_ga	0.003	1.003
blength_1	-0.139	0.870
blength_2	0.000	1.000
bw_1	-0.005	0.996
bw_2	2.552	12.835
inspired_oxygen.36_1	1.554	4.732
inspired_oxygen.36_2	1.487	4.423
weight_today.36_1	0.000	1.000
weight_today.36_2	-0.441	0.643
weight_today.36_3	0.000	1.000
disc_44	0.000	1.000
disc_44:ventilation_support_level_modified.441	0.282	1.326
disc_44:ventilation_support_level_modified.442	1.807	6.090
disc_44:p_delta.44	-0.001	0.999
disc_44:peep_cm_h2o_modified.44	0.028	1.028
disc_44:med_ph.441	1.006	2.736
disc_44:inspired_oxygen.44_1	0.223	1.250
disc_44:inspired_oxygen.44_2	-0.323	0.724
disc_44:weight_today.44_1	0.000	1.000
disc_44:weight_today.44_2	-0.447	0.640

4.3 Assessment of Model Fit

Figure 3 illustrate the discrimination and calibration of the model in the validation set. The ROC is above the diagonal line and very close to a rectangle shape, with an AUC very close to 1. This result suggest that under appropriate sensitivity-specificity trade-off, the model can discriminate positive/negative outcome successfully. On the calibration plot, the relationship between the predicted probabilities and observed proportions is roughly a straight line, with some variance when the predicted probability/observed probability is high. This suggest that the model predicts probability of Tracheotomy/Death well, with a small deviate when the predicted probability is high (false positive).

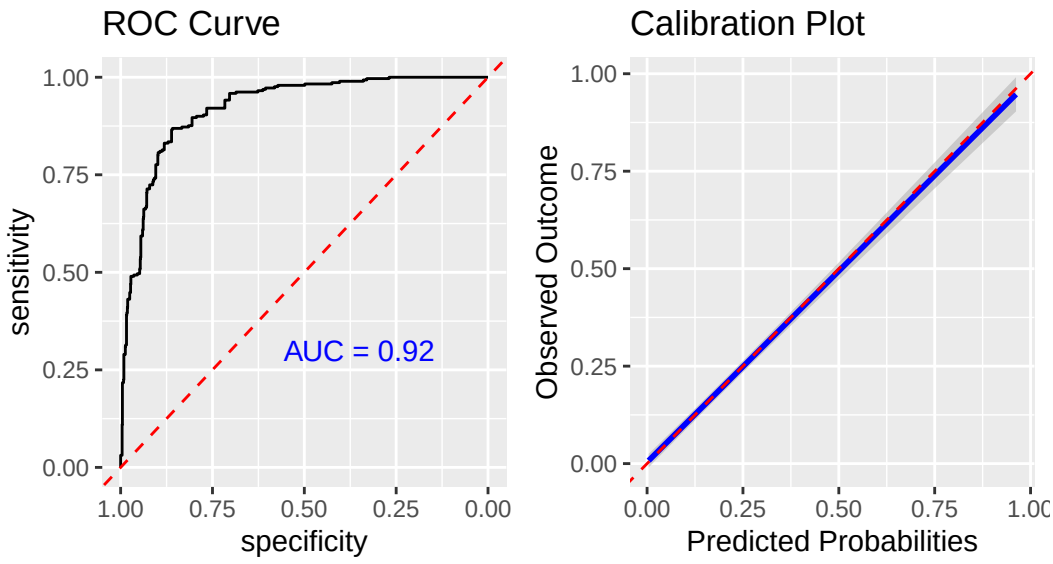

Figure 3: Model evaluation

5 Disscusion

In this analysis, we performed multiple imputations and lasso logistic regressions with cross-validation to build linear models to predict the single outcome of Tracheotomy and Death. Specifically, we conducted the imputation separately for individuals discharged before or after 44 week. We combine the imputation datasets and splited them into training set and validation set. We added an indicator term for those discharge before and after 44 week and form interaction terms for selected variables. We pooled the coeficients from the models with different imputation datasets for validation.

Our analysis have certain strength. The seperate-imputation and the model with the specific interaction terms can treat the patients discharged before or after 44 week differently, but can still incorporate both indiciduals in estimating coeficients not related to the 44 week. Thus, this method ensures a comprehensive evaluation of the impact of various covariates on the outcome, with due consideration for the discharge status at the 44th week.

On the other hand, pooling coefficients from the model estimated with different imputation datasets ensures stability and reliability in variable selection.

However, in this analysis, we didn't explore the multilinearity patterns in the covariates and thus didn't consider to include other interaction terms. Also, we didn't handle the outliers at certain covariates and the multip-centers effect.

6 Code Appendix

```
# Check and install packages
packages_to_check <- c("gtsummary", "gt", "tidyverse", "kableExtra", "mice",
                       "viridis", "knitr", "gridExtra",
                       "GGally", "pROC", "glmnet", "splines")
# Check if each package is installed and load it if available;
# otherwise, install and load it
for (pkg in packages_to_check) {
  if (!require(pkg, character.only = TRUE, quietly = TRUE)) {
    # If the package is not installed, install it
    install.packages(pkg)

    # Now, load the package
    library(pkg, character.only = TRUE)
  }
}

knitr::opts_chunk$set(warning = FALSE, message = FALSE,
                      echo = FALSE, fig.align = "center")
df <- read.csv("project2.csv")
# Data pre-processing
## Remove duplicated entries
df <- df[!duplicated(df$record_id), ]
df <- df[!df$center==21 | is.na(df$center),]
#df$com_prenat_ster[df$prenat_ster=="Yes"] <- "Yes"
#df <- select(df, -com_prenat_ster)
# Convert character variables to factor variables (for mice)
for (col in names(df)) {
  if (is.character(df[[col]])) {
    df[[col]] <- as.factor(df[[col]])
  }
}
#convert factor variables coded as numeric variables to the right format
fac_vars <- c("center","mat_race", "mat_ethn", "del_method",
              "ventilation_support_level.36", "med_ph.36",
              "ventilation_support_level_modified.44","med_ph.44")
for (var in fac_vars) {
  df[[var]] <- as.factor(df[[var]])
}
```

```r
# EDA for any_surf (which contain so many missing values)
# df %>%
#   mutate(missing_surf = ifelse(is.na(any_surf), 1, 0),
#          center = as.factor(center)) %>%
#   select(-c(record_id, any_surf)) %>%
#   tbl_summary(by = missing_surf,
#               percent = "row") %>%
#   add_p() %>%
#   modify_spanning_header(all_stat_cols() ~ "**If missing any_surf**")


## Separate into two dataset by discharge time
variables.44 <- names(df)[grep("\\.44$", names(df))]
df_clean <-
  df %>%
  mutate(trach_or_death = ifelse(Trach == 1 | Death == "Yes", 1, 0),
         if_any_record_44 = !is.na(variables.44[1]) |
              !is.na(variables.44[2])|
              !is.na(variables.44[3]) | !is.na(variables.44[4])|
              !is.na(variables.44[5])| !is.na(variables.44[6]),
         cohort = case_when(!if_any_record_44 & is.na(hosp_dc_ga) ~ "cohort 36",
                            hosp_dc_ga>=0 & hosp_dc_ga < 44 ~ "cohort 36",
                            hosp_dc_ga>= 44 ~ "cohort 44",
                            if_any_record_44 & is.na(hosp_dc_ga) ~"cohort 44")) %>%
  select(-c(any_surf, Trach, Death, if_any_record_44, com_prenat_ster)) %>%
  filter(center!=21 | is.na(center))

df_44 <-
df_clean %>%
  filter(cohort == "cohort 44") %>%
  select(-cohort)

df_36 <-
  df_clean %>%
  filter(cohort == "cohort 36") %>%
  select(-c(cohort, variables.44))
set.seed(56)
test_num_36 <- sample(c(TRUE, FALSE), size = nrow(df_36),
                      replace = TRUE, prob = c(0.3, 0.7))
test_num_44 <- sample(c(TRUE, FALSE), size = nrow(df_44),
                      replace = TRUE, prob = c(0.3, 0.7))
```

```r
### Caution!! It would take > 0.5h to run mice
# ## Mice
# df_36_mice_out <- mice(df_36, m = 5, ignore = test_num_36,
#                        printFlag = FALSE, seed = 222)
#
#
# df_44_mice_out <- mice(df_44, m = 5, ignore = test_num_44,
#                        printFlag = FALSE, seed = 222,
#                        nnet.MaxNWts = 4300) # nnet.MaxNWts: allow it to run longer
# saveRDS(df_36_mice_out, file = "df_36_mice_out.RDS")
# saveRDS(df_44_mice_out, file = "df_44_mice_out.RDS")
#Load in the mice objects
df_36_mice_out <- readRDS("df_36_mice_out.RDS")
df_44_mice_out <- readRDS("df_44_mice_out.RDS")

# Storing complete training set and testing sets
df_36_train <- vector("list",5)
for (i in 1:5){
   df_36_train[[i]] <- mice::complete(filter(df_36_mice_out, !test_num_36),i)
}
df_36_test <- mice::complete(filter(df_36_mice_out, test_num_36), action="stacked")


df_44_train <- vector("list",5)
for (i in 1:5){
   df_44_train[[i]] <- mice::complete(filter(df_44_mice_out, !test_num_44),i)
}
df_44_test <- mice::complete(filter(df_44_mice_out, test_num_44), action="stacked")


## Combine the imputations from the two cohorts
df_train <- vector("list", 5)
for (i in 1:5) {
  #filling zeros for NAs in Cohort 36 (these values is not used in the glm)
   for (j in variables.44) {
       df_36_train[[i]][[j]] <- 0
   }
  # add indicator variable for the two cohorts
  df_36_train[[i]]$disc_44 <- 0
  df_44_train[[i]]$disc_44 <- 1
  df_train[[i]] <- rbind(df_36_train[[i]], df_44_train[[i]]) %>%
```

```
    select(-c(record_id))
}

for (j in variables.44) {
       df_36_test[[j]] <- 0
}
df_36_test$disc_44 <- 0
df_44_test$disc_44 <- 1

df_test <- rbind(df_36_test, df_44_test)
## Missing value heatmap
missing_heatmap <- function(df, main=""){
  df_missing <-
  apply(df, 2, function(x) ifelse(is.na(x), "Yes", "No")) %>%
  as.data.frame()

df_missing %>%
  mutate(record_id = as.factor(df$record_id)) %>%
  pivot_longer(cols = names(df)[2:length(names(df))],
               names_to = "Variables",
               values_to = "is_missing") %>%
  mutate(is_missing = factor(is_missing, levels= c("Yes", "No")))%>%
  mutate(Variables = factor(Variables, levels = names(df)[2:length(names(df))])) %>%
  ggplot(aes(Variables, record_id, fill=is_missing, )) +
  geom_tile() +
  scale_fill_viridis(discrete = TRUE, option = "G") +
  theme(text = element_text(size = 7),
        axis.text.x = element_text(size = 4,
                                   angle = 60, hjust = 0.8, vjust = .9),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
  labs(fill = "Is missing?", title = main)
}
missing_heatmap(df)
# missing_36 <- missing_heatmap(df_36, main = "Cohort 36")
# missing_44 <- missing_heatmap(df_44, main = "Cohort 44")
# grid.arrange(missing_36, missing_44, ncol=2)
## Summarized table
theme_gtsummary_compact(font_size = 4)
df %>%
  mutate(center = as.factor(center),
```

```
        trach_or_death = ifelse(Trach == 1 | Death == "Yes", "Yes", "No")) %>%
  select(-record_id, Trach, Death) %>%
  mutate(mat_ethn = case_when(mat_ethn== 1 ~ "Hispanic or Latino",
                              mat_ethn == 2 ~ "Not Hispanic or Latino",
                              is.na(mat_ethn) ~ NA),
         del_method = case_when(del_method == 1 ~ "Vaginal delivery",
                                del_method == 2 ~ "Cesarean section",
                                is.na(del_method) ~ NA),
         ventilation_support_level.36 = case_when(ventilation_support_level.36
== 0 ~ "No respiratory support or supplemental oxygen",
ventilation_support_level.36 == 1 ~ "Non-invasive positive pressure",
ventilation_support_level.36 == 2 ~ "Invasive positive pressure",
is.na(ventilation_support_level.36) ~ NA),
          ventilation_support_level_modified.44 = case_when(ventilation_support_level_modi
== 0 ~ "No respiratory support or supplemental oxygen",
ventilation_support_level_modified.44 == 1 ~ "Non-invasive positive pressure",
ventilation_support_level_modified.44 == 2 ~ "Invasive positive pressure",
is.na(ventilation_support_level_modified.44) ~ NA),
         Trach = case_when(Trach == 0 ~ "No",
                           Trach == 1 ~ "Yes"),
          med_ph.36 = case_when(med_ph.36 == 0 ~ "No",
                                med_ph.36 == 1 ~ "Yes",
                                is.na(med_ph.36) ~ NA),
          med_ph.44 = case_when(med_ph.44 == 0 ~ "No",
                                med_ph.44 == 1 ~ "Yes",
                                is.na(med_ph.44) ~ NA)) %>%

  tbl_summary(type = list(prenat_ster ~ 'categorical',
                          com_prenat_ster ~ 'categorical',
                          mat_chorio ~ 'categorical',
                          any_surf ~ 'categorical',
                          med_ph.36 ~ 'categorical',
                          med_ph.44 ~ 'categorical',
                          Trach ~ 'categorical',
                          Death ~ 'categorical'),
              by = trach_or_death,
              percent = "row",
              missing_text = "missing",
              statistic = list(all_continuous() ~ "{mean} ({sd})"),
              label = list(`center` = "Medical Center",
                           `mat_race` = "Maternal Race",
```

```
                          `mat_ethn` = "Maternal Ethnicity",
                          `bw` = "Birth Weight (g)",
                          `ga` = "Obstetrical Gestational Age",
                          `blength` = "Birth Length (cm)",
                          `birth_hc` = "Birth Head Circumference (cm)",
                          `del_method` = "Delivery Method",
                          `prenat_ster` = "Prenatal Corticosteroids",
                          `com_prenat_ster` = "Complete Prenatal Steroids",
                          `mat_chorio` = "Maternal Chorioamnionitis",
                          `gender` = "Gender",
                          `sga` = "Small for Gestational Age",
                          `any_surf` = "Surfactant Received",
                          `weight_today.36` = "Weight at 36 Weeks",
                          `ventilation_support_level_modified.36` = "Ventilation Support
                          `inspired_oxygen.36` = "Inspired Oxygen at 36 Weeks",
                          `p_delta.36` = "Peak Inspiratory Pressure at 36 Weeks",
                          `peep_cm_h2o_modified.36` = "PEEP* at 36 Weeks",
                          `med_ph.36` = "Medication for PH* at 36 Weeks",
                          `weight_today.44` = "Weight at 44 Weeks",
                          `ventilation_support_level_modified.44` = "Ventilation Support
                          `inspired_oxygen.44` = "Inspired Oxygen at 44 Weeks",
                          `p_delta.44` = "Peak Pressure at 44 Weeks",
                          `peep_cm_h2o_modified.44` = "PEEP at 44 Weeks",
                          `med_ph.44` = "Medication for PH at 44 Weeks",
                          `hosp_dc_ga` = "Hospital Discharge Gestational Age",
                          `Trach` = "Tracheostomy")) %>%
  bold_labels() %>%
  add_p() %>%
  add_overall() %>%
  modify_header(label = "**Variables**") %>%
  modify_spanning_header(all_stat_cols() ~ "**Tracheostomy or Death**") %>%
  modify_footnote(all_stat_cols() ~ "PEEP: Positive end exploratory pressure;
                  PH: Pulmonary Hypertension")%>%
  as_kable_extra(booktabs = TRUE,
                 format = "latex",
                longtable = TRUE,
                  caption = "Summary of variables by outcome")  %>%
  kable_styling(latex_options = "hold_position", font_size = 4)
linearty_check <- function(var, label){
  min <- min(df_clean[[var]], na.rm = TRUE)
  max <- max(df_clean[[var]],na.rm = TRUE)
```

```
  bin <- (max - min)/15
  plot <-
  ggplot(df_clean, aes(x = cut(df_clean[[var]], breaks = seq(min, max, by = bin)),
                       y = trach_or_death)) +
  stat_summary(fun = "mean", geom = "bar") +
  labs(title = "",
       x = label,
       y = "") +
  theme_minimal()+
  theme(axis.text.x = element_text(size = 3,
                                   angle = 60, hjust = 0.8, vjust = .9),
        axis.text.y = element_text(size = 4),
        axis.title.x = element_text(size = 6))
  return(plot)
}
bw <- linearty_check(var = "bw", "Birth weight (g)")
ga <- linearty_check(var = "ga", "Obstetrical gestational age")
bl <- linearty_check(var = "blength", "Birth length (cm)")
bhc <- linearty_check(var = "birth_hc", "Birth head\n circumference (cm)")
w_36 <- linearty_check(var = "weight_today.36", "Weight at 36 weeks")
iox_36 <- linearty_check(var = "inspired_oxygen.36",
               "Fraction of Inspired Oxygen\n at 36 weeks")
pd_36 <- linearty_check(var = "p_delta.36",
               "Peak Inspiratory Pressure\n (cmH2O) at 36 weeks")
peep_36 <- linearty_check(var = "peep_cm_h2o_modified.36",
               "Positive and exploratory\n pressure (cm H2O) at 36\n weeks")

w_44 <- linearty_check(var = "weight_today.44", "Weight at 44 weeks")
iox_44 <- linearty_check(var = "inspired_oxygen.44",
               "Fraction of Inspired\n Oxygen at 44 weeks")
pd_44 <- linearty_check(var = "p_delta.44",
               "Peak Inspiratory Pressure\n (cmH2O) at 44 weeks")
peep_44 <- linearty_check(var = "peep_cm_h2o_modified.44",
               "Positive and exploratory\n pressure (cm H2O) at 44\n weeks")

hosp <- linearty_check(var = "hosp_dc_ga", "Hospital Discharge Gestational Age")

print(grid.arrange(bw, ga, bl, bhc, w_36, iox_36, pd_36,
                   peep_36,w_44, iox_44, pd_44, peep_44, hosp,
                   ncol=4,
                   left = "Proportion of Trach or Death"))
```

```r
create_and_replace_spline <- function(imputed_data) {
  # Create spline term
  bw_spline <- bs(imputed_data$bw, knots = 1610,
                                degree = 1, intercept = FALSE)
  blength_spline <- bs(imputed_data$blength, knots = 40,
                                degree = 1, intercept = FALSE)
  inspired_oxygen_36_spline <- bs(imputed_data$inspired_oxygen.36, knots = .789,
                                degree = 1, intercept = FALSE)
  inspired_oxygen_44_spline <- bs(imputed_data$inspired_oxygen.44, knots = .737,
                                degree = 1, intercept = FALSE)
  weight_today_36_spline <- bs(imputed_data$weight_today.36, knots = c(1510,2710),
                                degree = 1, intercept = FALSE)
    weight_today_44_spline <- bs(imputed_data$weight_today.44, knots = 2110,
                                degree = 1, intercept = FALSE)

  # Replace original variable
  imputed_data <- imputed_data %>% select(-c(bw,blength,
                                             inspired_oxygen.36,
                                             inspired_oxygen.44,
                                             weight_today.36,
                                             weight_today.44))
  imputed_data$blength_1 <- blength_spline[,1]
  imputed_data$blength_2 <- blength_spline[,2]

  imputed_data$bw_1 <- bw_spline[,1]
  imputed_data$bw_2 <- bw_spline[,2]

  imputed_data$inspired_oxygen.36_1 <- inspired_oxygen_36_spline[,1]
  imputed_data$inspired_oxygen.36_2 <- inspired_oxygen_36_spline[,2]

  imputed_data$inspired_oxygen.44_1 <- inspired_oxygen_44_spline[,1]
  imputed_data$inspired_oxygen.44_2 <- inspired_oxygen_44_spline[,2]

  imputed_data$weight_today.36_1 <- weight_today_36_spline[,1]
  imputed_data$weight_today.36_2 <- weight_today_36_spline[,2]
  imputed_data$weight_today.36_3 <- weight_today_36_spline[,3]

  imputed_data$weight_today.44_1 <- weight_today_44_spline[,1]
  imputed_data$weight_today.44_2 <- weight_today_44_spline[,2]
  return(imputed_data)
}
```

```r
df_train_with_spline <- lapply(df_train, create_and_replace_spline)
df_test_with_spline <- create_and_replace_spline(df_test)

variables_other <- names(df_train_with_spline[[1]])[c(1:15, 22:27, 30:32)]
variables.44 <- names(df_train_with_spline[[1]])[c(17:20, 28:29, 33:34)]


model_formula <- as.formula(paste(
  "trach_or_death ~ ",
  paste0(variables_other[!variables_other %in% c("inspired_oxygen.36", "inspired_oxygen.44
  " + disc_44 +  disc_44 : (",
  paste0(variables.44, collapse = " + "), ")"
))


mod <- glm(data = df_train_with_spline[[1]],
            model_formula,
            family = "binomial")
##########################################################
#### Lasso ####
##########################################################
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(model_formula, data = df)[,-1]#Dropping intercept term since lasso
  y.ord <- df$trach_or_death

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

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

```r
  # Get coefficients
  coef <- coef(lasso_mod_cv, s = "lambda.min")
  return(coef)
}
# Find average lasso coefficients over imputed datasets
pooling_lasso <- function(traing_data){
  lasso_coef1 <- lasso(traing_data[[1]])
  lasso_coef2 <- lasso(traing_data[[2]])
  lasso_coef3 <- lasso(traing_data[[3]])
  lasso_coef4 <- lasso(traing_data[[4]])
  lasso_coef5 <- lasso(traing_data[[5]])
  lasso_coef <- cbind(lasso_coef1, lasso_coef2, lasso_coef3,
                      lasso_coef4, lasso_coef5)
  avg_coefs_lasso <- apply(lasso_coef, 1, mean)
  return(avg_coefs_lasso)
}
set.seed(6)
coef_pred <- pooling_lasso(df_train_with_spline)
# Create a data frame for odds ratios and their CIs
# Extract odds ratios and their confidence intervals
result_table <-
data.frame(
  Lasso_coefficient = round(coef_pred, 3),
  Odds_ratio = round(exp(coef_pred), 3)
  )

rownames(result_table)[rownames(result_table) == "(Intercept)"] <- "Intercept"
result_table %>%
  kableExtra::kable(booktabs = TRUE,
                    longtable = TRUE,
                    col.names = c("Lasso coefficient", "Odds ratio"),
  caption = "Association of selected variables with tracheostomy and death")%>%
  kable_styling(latex_options = "hold_position", font_size = 4)
# Get predicted values with pooled coefs
predict_lasso <- function(test_data, coef){
  x_vars <- model.matrix(model_formula, test_data)
  test_data$lasso <- x_vars %*% coef
  mod_lasso <- glm(trach_or_death~lasso, data = test_data, family = "binomial")
  predict_probs_lasso <- predict(mod_lasso, type="response")
  return(predict_probs_lasso)
}
```

```r
df_test_with_spline$predict_probs_lasso <- predict_lasso(df_test_with_spline, coef = coef_


#ROC-AUC
roc_curve_lasso <- roc(response = df_test_with_spline$trach_or_death,
                       predictor = df_test_with_spline$predict_probs_lasso,
                       levels = c(0,1), direction= "<")
auc_value <- auc(roc_curve_lasso)

roc_plot <- ggroc(roc_curve_lasso) +
  ggtitle("ROC Curve")+
  geom_abline(intercept = 1, slope = 1, linetype = "dashed", color = "red") +
  annotate("text", x = 0.35, y = 0.3,
           label = paste("AUC =", round(auc_value, 2)), color = "blue")
  theme_minimal()

# Calibration
cal_plot <- ggplot(df_test_with_spline, aes(predict_probs_lasso, trach_or_death)) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(title = "Calibration Plot",
       x = "Predicted Probabilities",
       y = "Observed Outcome")

print(grid.arrange(roc_plot, cal_plot, ncol=2))
```