Project 2

Destiny Rankins

Abstract

Introduction

An estimated 10000 to 15000 newborns develop Bronchopulmonary Dysplasia (BPD) in the United States every year [1]. BPD is a form of chronic lung disease that affects newborns, most often those who are born prematurely and need oxygen therapy [1]. BDP is the result of a newborn's lungs not developing normally while the baby is growing in womb, or not developing fully if the baby was born premature [2]. The lungs develop around 30-40 weeks. While the condition is a consequence of being born extremely premature, the severity of the disease varies from infant to infant. There is no particular cure for Bronchiopulmonary Dysplasia however, there are various treatment options to help support infant's lungs, allowing for them to heal and grow [1]. Some of these treatment options include diuretics, bronchiodilators, corticosteroids, viral immunization, and cardiac medications.

There are varying severity levels used to classify infants with BPD; mild (grade 1), moderate (grade 2), severe (grade 3), and very severe (grade 4). Infants with Grade 2 BPD are defined as those with the need for non-invasive positive pressure at 36 weeks post-menstrual age (PMA) [3]. The most severely affected infants (grade 3 BPD) are dependent on a ventilator at 36 week PMA moreover, they have a need for ongoing invasive positive pressure ventilation (IPPV) [4]. Infants with severe BPD must be discharged from the hospital on a ventilator. This requires a tracheostomy, a procedure to help air and oxygen reach the lungs by creating an opening into the trachea (windpipe) from outside the neck [5]. The surgical opening in the neck allows them to be connected to a ventilator. Up to 12% of infants with severe or grade 3 BPD require a trachesotomy. Some potential benefits of performing tracheostomy include providing a stable airway, improving growth, promoting age-appropriate interactions, and improving participation in developmental care.

Exploratory Data Analysis

There are a total of thirty categorical and continuous variables in the data set. There are categorical variables in the data set for: the medical center, the mother's race, the mother's ethnicity, the birthing delivery method (vaginal delivery or Cesarean section), prenatal corticosteriods, completed prenatal steroids, maternal chorioamnionitis, infant gender, whether infant was small for gestational age, if the infant received surfactant at any point in the first 72 hours, level of ventilation support at 36 and 44 weeks (0 = no respiratory support or supplemental oxygen, 1 = non-invasive positive pressure, and 2 = invasive positive pressure), if medication for pulmonary hypertension was received at 36 and 44 weeks, if the infant received a tracheostomoy at discharge, and death before discharge. There are continuous variables in the data set for: the birth weight in grams, the obstetrical gestational age, the birth length in centimeters, the birth head circumference in centimeters, the infant weight at 36 and 44 weeks, the fraction of inspired oxygen at 36 and 44 weeks, the peak inspiration pressure (cmH20) at 36 and 44 weeks, positive and exploratory pressure (cmH20) at 36 and 44 weeks, and hospital discharge gestational age

When observing the demographic statistics from the data in Table 1, we see that most of the participants are non-Hispanic or Latino. Moreover, we see that 59% of the non-Hispanic or Latino maternal participants selected “Other” for race and 33% selected American Indian or Alaskan Native. In contrast, 65% of Hispanic maternal participants selected “Asian” for race and 32% is unknown. Since we do not know for sure if an error occurred when creating the code book where “6” is meant to specify “Other” for race or if there was a clerical data entry error in the data set where “0” is a specifier for race, these values were treated as unknown. None of the participants identified as Black or African American, Native Hawaiian or Other Pacific Islander or, White. From Table 2, we observe that there were more males (466) than females (315) that were small for gestational age. The average birth weight was higher for males (838 grams) than females (762 grams). Furthermore, the average weight for infant males for males at 36 weeks PMA and 44 weeks PMA than for females. In Table 3: Summary of Respiratory Support Variables, we observe the overall average for infants that did not receive respiratory support or supplemental oxygen increase from 12% to 47%. The overall average of number of infants that needed non-invasive positive pressure decreased from 61% to 25% for 36 weeks PMA and 44 week PMA, respectively. An average of 260 infants had invasive support for ventilation at 36 weeks PMA and an average of 157 infants had invasive support for ventilation at 44 weeks PMA. From Table 4, we see that most of the participants took prenatal steroids (87%), while we observe an 11% decrease for participants that completed taking the prenatal steroids.In this table, we also see that an average of 145 infants (15%) had a tracheostomy at discharge and an average of 54 infants (5.4%) died before discharge.

Overall, 13.8% of the data is missing. Furthermore, there are seven variables with more than 40% of missing data. It appears to be a significant amount of infant data that is missing at 44 weeks post-menstrual age. These include the variables for the fraction of inspired oxygen at 44 weeks, the peak inspiratory pressure (cm H20) needed at 44 weeks, weight at 44 weeks, the positive and exploratory pressure (cm H20) needed at 44 weeks, whether the infant received surfactant at any point in the first 72 hours, the ventilation support level at 44 weeks, and whether medication for pulmonary hypertension at 44 weeks was received. This could be the case if the invasive ventilation support was remove before 44 weeks post-mentstrual age. After grouping the missing data by the variable for medical center, it appears that a lot of the missing

Table 1: **Table 1. Average Birth and Infants Weights**

Variable	**Overall**, N = 992	**Female**, N = 408	**Male**, N = 584
___Average Birth Weight {g}___			
Mean	807	762	839
___Average Weight at 36 Weeks {g}___			
Mean	2,122	2,028	2,188
Missing	90	34	56
___Average Weight at 44 Weeks {g}___			
Mean	3,648	3,530	3,730
Missing	446	185	261

Table 2: **Table 2. Maternal Demographics**

	Overall, N = 939	**Hispanic or Latino**, N = 74	**Not Hispa
___Race of Mother___			
American Indian or Alaskan Native	286 (31%)	2 (2.9%)	
Asian	111 (12%)	44 (65%)	
Unknown	534 (57%)	22 (32%)	
Missing	8	6	

Table 3: **Table 4. Summary of Steroid Status and Chorioamnionitis**

	Overall, N = 992	**Female**, N = 408	**Male**, N = 584
___Average Tracheostomy___			
0	846 (85%)	348 (85%)	498 (85%)
1	146 (15%)	60 (15%)	86 (15%)
___Average Death___	54 (5.5%)	17 (4.2%)	37 (6.3%)
Missing	2	1	1
___Prenatal Steroids___	831 (87%)	334 (84%)	497 (89%)
Missing	35	11	24
___Completed Prenatal Steroids___	606 (76%)	244 (77%)	362 (75%)
Missing	193	90	103
___Maternal Chorioamnionitis___	160 (17%)	65 (17%)	95 (17%)
Missing	62	23	39

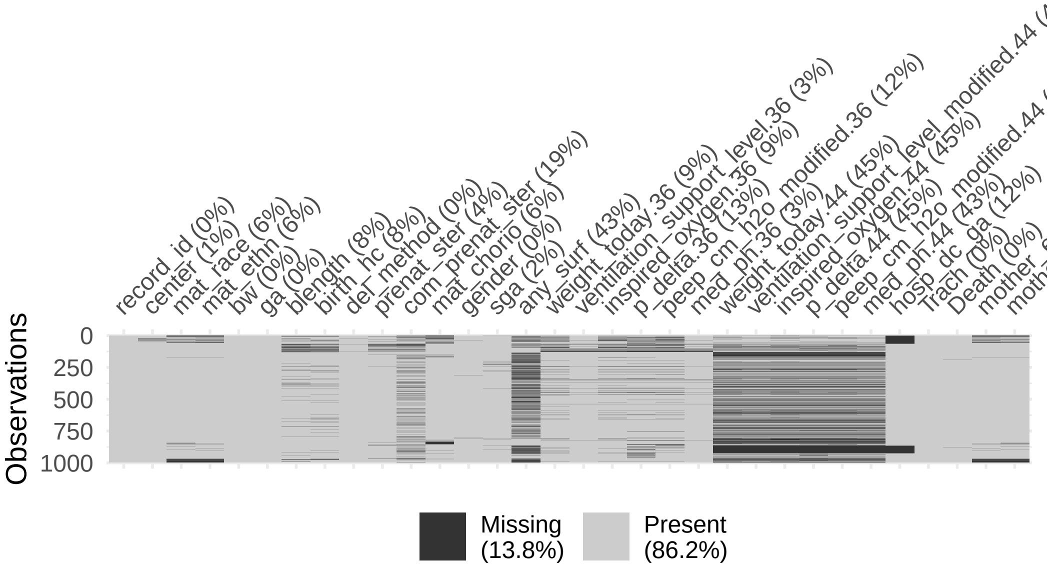
Table 4: **Table 3. Summary of Respiratory Support Variables**

	Overall, N = 992	**Female**
___Ventilation Support Level at 36 Weeks PMA___		
0	116 (12%)	52 (13
1	586 (61%)	240 (6
2	260 (27%)	101 (2
Missing	30	15
___Ventilation Support Level at 44 Weeks PMA___		
0	267 (47%)	113 (4
1	144 (25%)	58 (25
2	157 (28%)	60 (26
Missing	424	177
___Fraction of Inspired Oxygen at 36 Weeks PMA___		
Mean	0.34	0.3
Missing	91	37
___Fraction of Inspired Oxygen at 44 Weeks PMA___		
Mean	0.34	0.3
Missing	448	187
___Peak Inspiratory Pressure (cmH2O) at 36 weeks PMA___		
Mean	5	5
Missing	128	57
___Peak Inspiratory Pressure (cmH2O) at 44 weeks PMA___		
Mean	8	7
Missing	448	187
___Positive and Exploratory Pressure (cm H2O) at 36 weeks___		
Mean	6	6
Missing	117	50
___Positive and Exploratory Pressure (cm H2O) at 36 Weeks PMA___		
Mean	4	4
Missing	446	185
___Medication for Pulmonary Hypertension at 36 weeks___		
0	896 (93%)	368 (9
1	66 (6.9%)	25 (6.
Missing	30	15
___Medication for Pulmonary Hypertension at 44 weeks___		
0	469 (83%)	184 (8
1	99 (17%)	47 (20
Missing	424	177

Table 5: Variables with Significant Missingness

variable	n_miss	pct_miss
inspired_oxygen.44	448	44.980
p_delta.44	448	44.980
weight_today.44	446	44.779
peep_cm_h2o_modified.44	446	44.779
any_surf	433	43.474
ventilation_support_level_modified.44	424	42.570
med_ph.44	424	42.570

data at 44 weeks is from medical center 2. There is also a significant amount of missing data for the variable any_surf (did the infant receive surfactant at any point in the first 72 hours) and 46% is from medical center 2. This may be the case because practices and procedures can differ between medical centers i.e how different medical centers record data or do medical charting. Based on this assumption, the data could be considered missing not at random where the probability of missing data is systematically related to the hypothetical values that are missing[6].Some medical centers are referral centers, special centers, or academic tertiary centers.

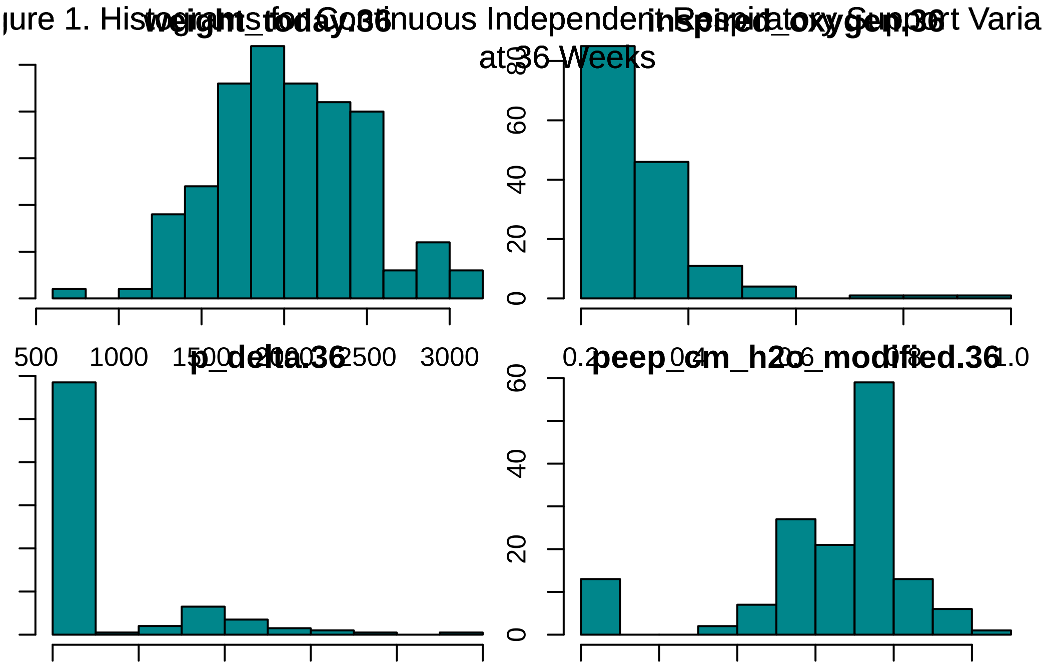


In Figure 1 and Figure 2, we observe the distribution for the continous independent respiratory variables at 36 and 44 weeks. The distributions for most of these variable appear to be random where we observe no apparant pattern. The distribution for the variable for fraction of inspired

Table 6: Medical Center 2 Missing Variable Summary

center	variable	n_miss	pct_miss
2	any_surf	295	46.825
2	inspired_oxygen.44	253	40.159
2	weight_today.44	252	40.000
2	peep_cm_h2o_modified.44	250	39.683
2	p_delta.44	248	39.365
2	ventilation_support_level_modified.44	239	37.937
2	med_ph.44	239	37.937
2	com_prenat_ster	105	16.667
2	peep_cm_h2o_modified.36	41	6.508
2	p_delta.36	39	6.190
2	weight_today.36	36	5.714
2	inspired_oxygen.36	36	5.714
2	birth_hc	29	4.603
2	blength	24	3.810

oxygen at 36 weeks appears to be right-skewed or positively skewed. The distribution for the variable for the weight at 44 weeks appears to be normal with a bell-shape. In Figure 2, we observe the distribution for the continuous independent birth variables. The distributions for the birth weight, birth height, and gestational age appear to be right-skewed.



ıure 2. Histograms for Continuous Independent Respiratory Support Variab
at 44 Weeks

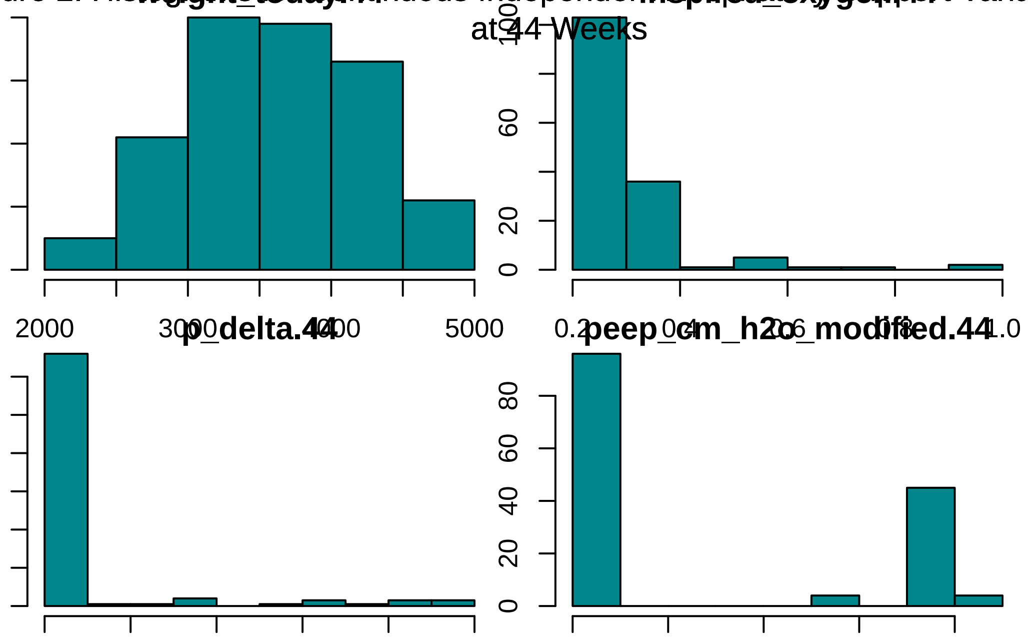
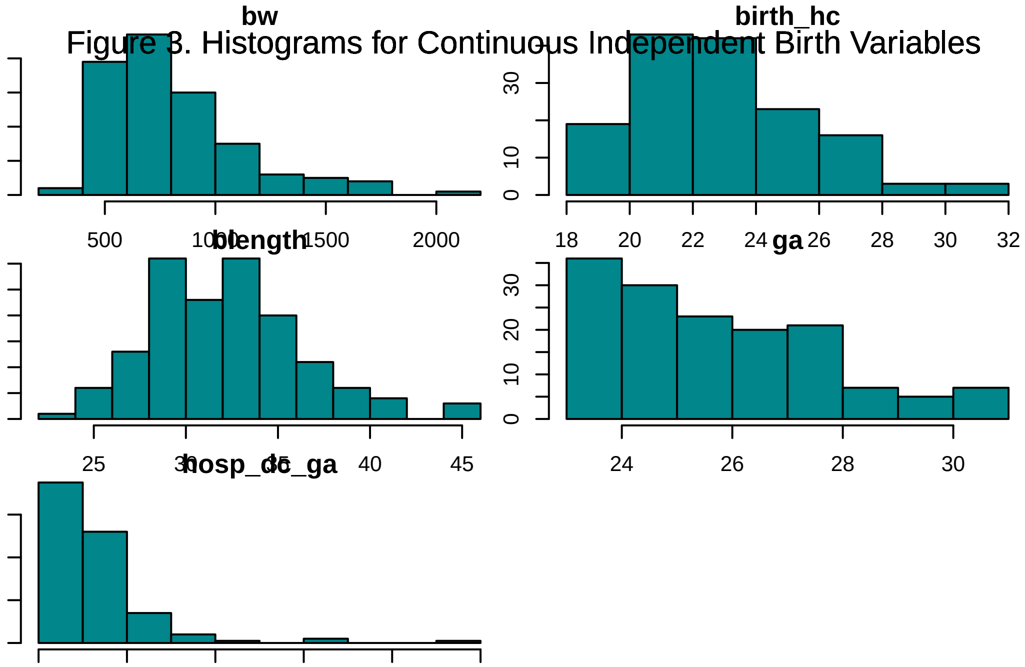
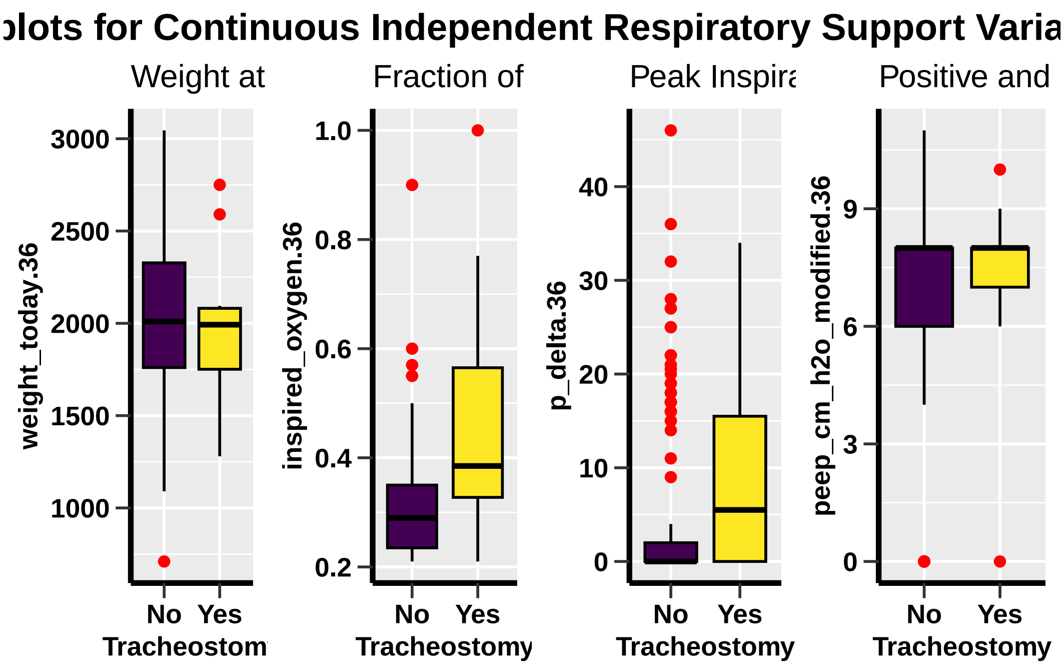
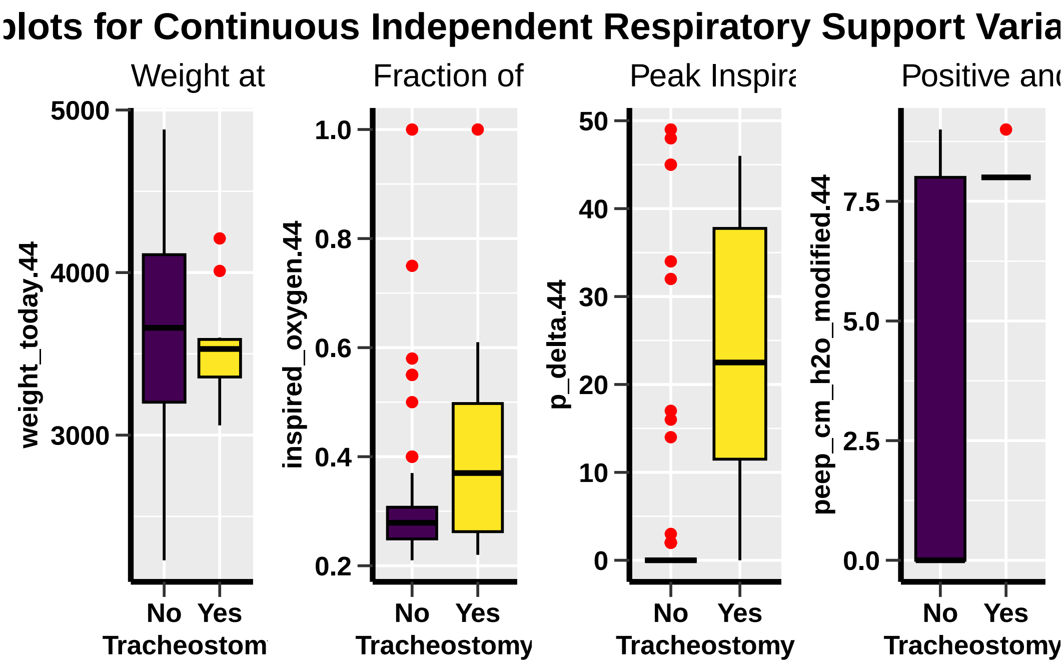
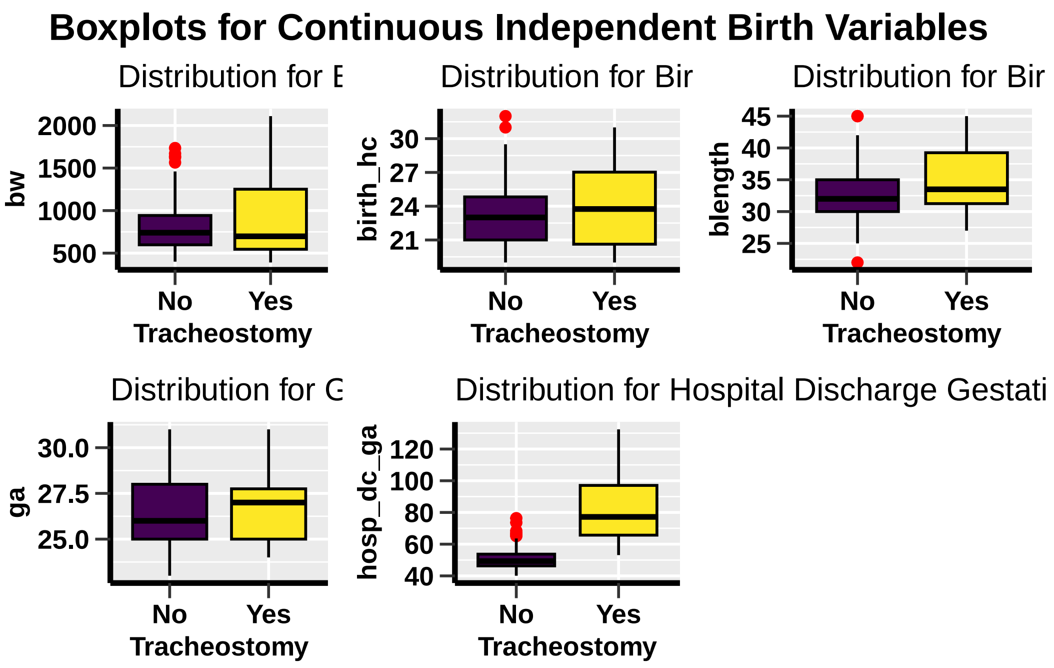


Figure 3. Histograms for Continuous Independent Birth Variables




plots for Continuous Independent Respiratory Support Varia
Weight at
Fraction of
Peak Inspira
Positive and
weight_today.36
3000
2500
2000
1500
1000
inspired_oxygen.36
1.0
0.8
0.6
0.4
0.2
p_delta.36
40
30
20
10
0
peep_cm_h2o_modified.36
9
6
3
0
No
Yes
Tracheostom
Tracheostomy


plots for Continuous Independent Respiratory Support Varia
Weight at
Fraction of
Peak Inspira
Positive and
weight_today.44
5000
4000
3000
inspired_oxygen.44
1.0
0.8
0.6
0.4
0.2
p_delta.44
50
40
30
20
10
0
peep_cm_h2o_modified.44
7.5
5.0
2.5
0.0
No
Yes
Tracheostom
Tracheostomy

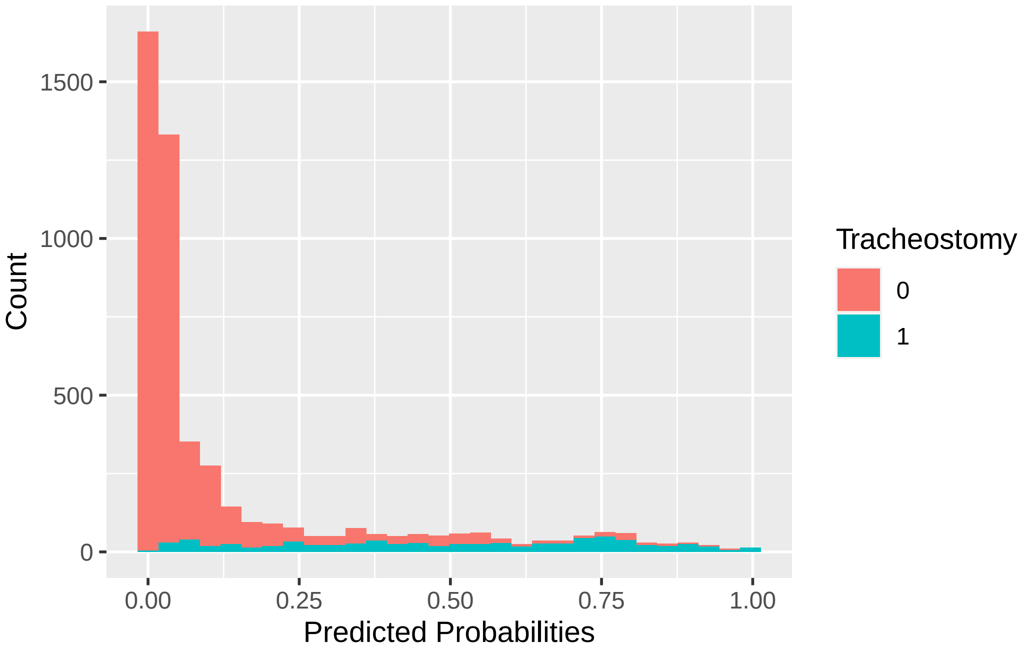

Boxplots for Continuous Independent Birth Variables

Methods

Model Development

Three logistic regression models were developed from twenty-six chosen categorical and continuous variables. The categorical variables were converted from character-type variables to factors, allowing for statistical analysis. For imputation, the data was to include “relevant” variables. To handle the missing data, the multiple imputation method using the mice() function was employed to generate five imputed data sets with filled missing values. Each imputed data set was stored in a list to be acessed when fitting the regression models. The logistic model was fit using the glm() function from the stats package where the model of binary outcomes with “family =”“binomial” was indicated. The Lasso regression model was fit using the cv.glmnet() function from the glmnet package. Ten-fold cross-validation was implemented by setting “nfolds = 10”, “foldid = folds” and, indicating model of binary outcomes with “family =”binomial”“. The Ridge model followed the same model-building procedure as described in the Lasso regression. However, setting”alpha = 1”, specifies the Lasso method and setting “alpha = 0” in the cv.glmnet() function indicates Ridge regression. This approach helps to identify the coefficients associated with the optimal   for each imputed set. The average of all coefficients derived from these five imputed sets to were used to predict the outcome of a tracheostomy.

Model Selection and Cross Validation



Model Performance

Models were evaluated using discrimination to assess the model's ability to differentiate between positive and negative outcomes. This evaluation included plotting the receiving operating characteristic curve (ROC), calculating the sensitivity, specificity, positive predicted values, negative predicted values and, overall accuracy for predicting tracheostomy placement. The models were calibrated using predicted outcome probabilities to help assess reliability or, how well the predicted probabilities of an event match the observed outcome. The predicted outcome probabilities were obtained using the predict() function. Brier score performance was used to measure the accuracy of the predicted probabilities. This measure further helps to assess how well each model's predicted outcome aligns with the actual observed values.

From Table 6: Regression Coeffiecient Comparison, we observe that the values for coefficients in logistic regression are generally larger than those in lasso and ridge regressions. This discrepancy arises because the regularization terms in lasso and ridge penalize large coefficients. From Figure : Model Performance, the AUC-ROC plot is used to visualize and determine the model's ability to discriminate differentiate between tracheostomy placement and no tracheostomy placement. A higher AUC score is associated with better discrimination. Here, the logistic model had the highest AUC score (0.933). From Figure , the logistic and lasso models seem to have more points aligned the perfect calibration diagonal line compared to ridge however, the logistic appears to have a better model calibration. From Table 8: Brier Score Measures, we observed the logistic model achieved lowest Brier Score (0.065). The lower the Brier Score, the better the model's performance.

Table 7: Regression Coefficient Comparisons

	coefs_logistic	coefs_lasso	coefs_ridge
(Intercept)	-4.960	-2.513	-4.233
center2	-0.581	-0.011	-0.242
center3	-6.750	-0.250	-0.442
center4	-1.430	-0.199	-0.152
center5	0.074	0.101	0.067
center7	-1.850	-0.015	-0.344
center12	1.635	0.799	0.981
center16	-1.666	0.000	-0.329
center20	-13.407	0.000	-0.454
center21	16.992	1.058	2.274
bw	0.001	0.000	0.000
ga	-0.196	0.000	0.007
blength	0.021	0.000	0.003
birth_hc	0.093	0.003	0.012
del_method2	0.496	0.051	0.129
prenat_sterYes	1.613	0.385	0.348
com_prenat_sterYes	0.355	0.040	0.151
mat_chorioYes	-0.356	0.000	0.026
genderMale	-0.213	0.000	-0.021
sgaSGA	0.202	0.000	0.020
any_surfYes	-0.124	0.000	0.050
weight_today.36	-0.001	0.000	0.000
ventilation_support_level.361	-0.415	-0.011	-0.252
ventilation_support_level.362	1.042	0.265	0.387
inspired_oxygen.36	2.886	0.601	1.059
p_delta.36	-0.038	0.000	0.006
peep_cm_h2o_modified.36	0.055	0.010	0.031
med_ph.361	-0.360	0.000	0.106
weight_today.44	0.000	0.000	0.000
ventilation_support_level_modified.441	0.022	0.000	-0.095
ventilation_support_level_modified.442	1.919	0.537	0.550
inspired_oxygen.44	-1.839	-0.068	0.253
p_delta.44	-0.008	0.000	0.010
peep_cm_h2o_modified.44	0.078	0.027	0.048
med_ph.441	0.772	0.178	0.309
hosp_dc_ga	0.028	0.008	0.007

Table 8: Model Performance Measures

Model	Sensitivity	Specificity	PPV	NPV	Accuracy
Logistic	0.886	0.873	0.545	0.978	0.875
Lasso	0.832	0.881	0.546	0.968	0.874
Ridge	0.870	0.840	0.484	0.974	0.845

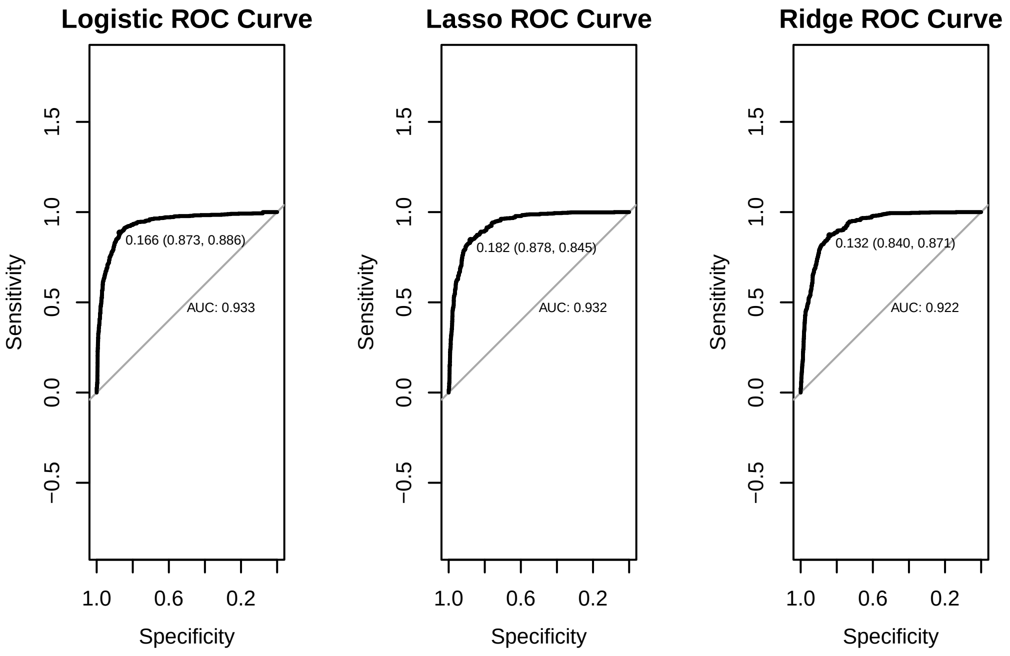
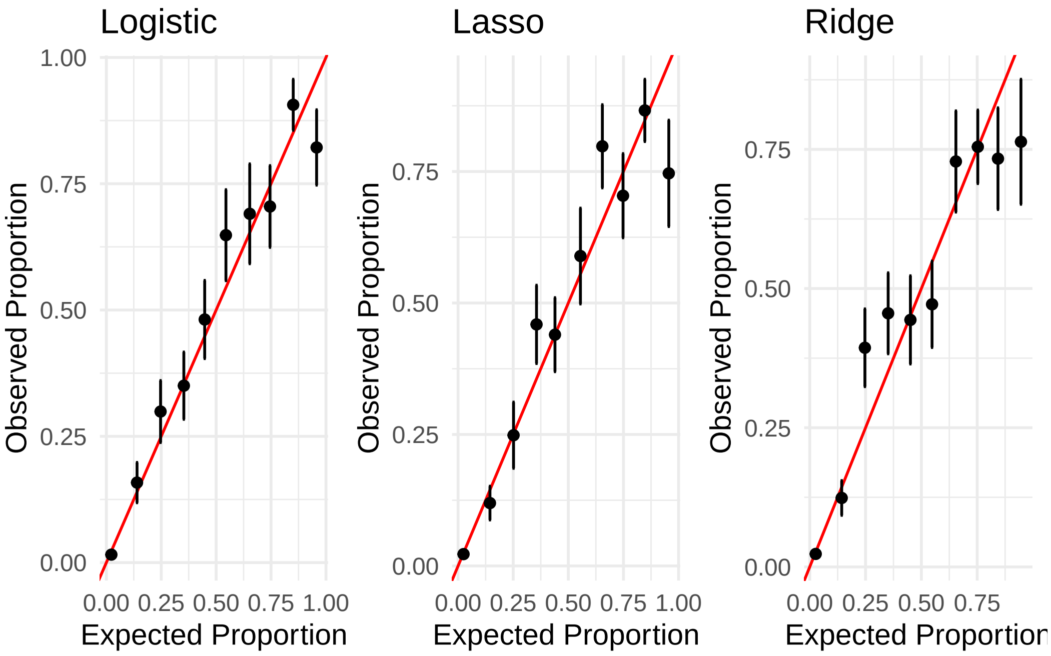


Table 9: Brier Score Measures

Models	BrierScore
Logistic	0.065
Lasso	0.069
Ridge	0.075



Discussion

The results of this regression analysis provide insights into the prediction of tracheostomy among infants diagnosed with severe bronchopulmonary dysplasia (BPD). Given the vulnerability of the patient population and the risk associated with a tracheostomy, it is important to develop a model that can predict the potential need for a invasive ventilation support. After fitting the logistic, lasso, and ridge models on the average coefficients from the imputed data sets, the logistic model appears to perform the best. We were able to assess the reliability and accuracy of these models using discrimination techniques and calibration methods. The current study has some limitations. There were very few observed outcomes for tracheostomy and death. Furthermore, we did not focus on the death outcome in this study it is uncertain that that cause of death is with complications from the tracheostomy for every infant. Significant missingness for predictors of interest was another limitation in the study. The multiple imputation method was used to impute missing variables. In this study, we did not split the data into a training set and test set for validation. Some of the regression model functions included cross-validation and we did implement however, it is helpful to have a independent test set to evaluate the final model's performance on unseen data. Furthermore, we did not conduct external validation on another data set to evaluate the generalizability of the model or the model's ability to be applied outside the clinical setting.

offering valuable insights into the complex clinical decision-making process for this vulnerable patient population.

Conclusion

References

1. Association, A. L. (n.d.). *Learn About Bronchopulmonary Dysplasia*. Www.lung.org. https://www.lung.org/lung-health-diseases/lung-disease-lookup/bronchopulmonary-dysplasia/learn-about-bpd#:~:text=Key%20Facts

2. *Newborn Breathing Conditions - Bronchopulmonary Dysplasia (BPD) | NHLBI, NIH*. (n.d.). Www.nhlbi.nih.gov. https://www.nhlbi.nih.gov/health/bronchopulmonary-dysplasia

3. Milenka Cuevas Guamán, Nikou Pishevar, Abman, S. H., Keszler, M., Truog, W. E., Panitch, H. B., & Nelin, L. D. (2021). Invasive mechanical ventilation at 36 weeks post-menstrual age, adverse outcomes with a comparison of recent definitions of bronchopulmonary dysplasia. *Journal of Perinatology*, *41*(8), 1936–1942. https://doi.org/10.1038/s41372-021-01102-w

4. Miller, A. N., Shepherd, E. G., Manning, A., Shamim, H., Chiang, T., El-Ferzli, G. T., & Nelin, L. D. (2023). Tracheostomy in Severe Bronchopulmonary Dysplasia—How to Decide in the Absence of Evidence. *Biomedicines*, *11*(9), 2572–2572. https://doi.org/10.3390/biomedicines11092572

5. *Tracheostomy*. (2023, April 11). Www.hopkinsmedicine.org. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/tracheostomy#:~:text=Tracheostomy%20is%20a%20procedure%20to

6. Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of School Psychology, 48(1), 5–37. https://doi.org/10.1016/j.jsp.2009.10.001

Code Appendix

```
# load libraries
suppressPackageStartupMessages(library(gtsummary))
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(naniar))
suppressPackageStartupMessages(library(knitr))
suppressPackageStartupMessages(library(kableExtra))
suppressPackageStartupMessages(library(tinytex))
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(mice))
suppressPackageStartupMessages(library(rmarkdown))
suppressPackageStartupMessages(library(corrplot))
suppressPackageStartupMessages(library(ggcorrplot))
suppressPackageStartupMessages(library(gridExtra))
suppressPackageStartupMessages(library(ggpubr))
suppressPackageStartupMessages(library(reshape2))
suppressPackageStartupMessages(library(leaps))
suppressPackageStartupMessages(library(glmnet))
suppressPackageStartupMessages(library(caret))
suppressPackageStartupMessages(library(stats))
suppressPackageStartupMessages(library(pROC))
suppressPackageStartupMessages(library(DescTools))
# load in the data set
project2 <- read.csv("C:/Users/desti/Downloads/project2.csv", header=TRUE)

# check for duplicates
duplicate_id <- duplicated(project2$record_id)

# remove duplicates
trach_df <- project2[!duplicate_id, ]

# change the variables to factors
trach_df[,c(2:4, 9:15, 17, 21, 23, 27, 29)] <- lapply(trach_df[,c(2:4, 9:15, 17, 21, 23, 2
# average birth weight, average birth weight at 36 week, and average weight at 44 weeks
trach_df %>%
  dplyr::select(gender, bw, weight_today.36, weight_today.44) %>%
    tbl_summary(
      by = gender,
      type = all_continuous() ~ "continuous2",
      statistic = all_continuous() ~ c("{mean}"),
      missing_text = "Missing",
      label = list(bw = "Average Birth Weight {g}",
                  weight_today.36 = "Average Weight at 36 Weeks {g}",
                  weight_today.44 = "Average Weight at 44 Weeks {g}")) %>%
      add_overall() %>%
      modify_header(label ~ "**Variable**") %>%
      modify_caption("**Table 1. Average Birth and Infants Weights**") %>%
      modify_spanning_header(c("stat_1", "stat_2") ~ "**Gender**") %>%
      bold_labels()%>%
    tbl_butcher()


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


# table displaying the summary of the mother's race and ethnicity
trach_df %>%
  dplyr::select(mother_race, mother_ethn) %>%
  tbl_summary(
    by = mother_ethn, missing_text = "Missing",
    label = list(mother_race = "Race of Mother")) %>%
  add_overall() %>%
  modify_header(label ~ " ") %>%
  modify_caption("**Table 2. Maternal Demographics**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Ethnicity of Mother**") %>%
  bold_labels() %>%
  tbl_butcher()

# average tracheostomy at discharge and death before discharge
trach_df %>%
  dplyr::select(Trach, Death, gender, prenat_ster, com_prenat_ster, mat_chorio) %>%
  tbl_summary(
    by = gender,
    type = all_continuous() ~ "continuous2",
    statistic = all_continuous() ~ c("{median}"),
    missing_text = "Missing",
    label = list(prenat_ster = "Prenatal Steroids",
                 com_prenat_ster = "Completed Prenatal Steroids",
                 mat_chorio = "Maternal Chorioamnionitis",
                 Trach = "Average Tracheostomy",
                 Death = "Average Death")) %>%
  add_overall() %>%
  modify_header(label ~ " ") %>%
  modify_caption("**Table 4. Summary of Steroid Status and Chorioamnionitis**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**Gender**") %>%
  bold_labels()%>%
  tbl_butcher()

# average tracheostomy at discharge and death before discharge
trach_df %>%
  dplyr::select(gender, ventilation_support_level.36, ventilation_support_level_modified.4
  tbl_summary(
    by = gender,
    type = all_continuous() ~ "continuous2",
    statistic = all_continuous() ~ c("{mean}"),
    missing_text = "Missing",
    label = list(ventilation_support_level.36 = "Ventilation Support Level at 36 Weeks PMA
                 ventilation_support_level_modified.44 = "Ventilation Support Level at 44
                 inspired_oxygen.36 = "Fraction of Inspired Oxygen at 36 Weeks PMA",
                 inspired_oxygen.44 = "Fraction of Inspired Oxygen at 44 Weeks PMA",
                 p_delta.36 = "Peak Inspiratory Pressure (cmH2O) at 36 weeks PMA",
                 p_delta.44 = "Peak Inspiratory Pressure (cmH2O) at 44 weeks PMA",
                 peep_cm_h2o_modified.36 = "Positive and Exploratory Pressure (cm H2O) at
                 peep_cm_h2o_modified.44 ="Positive and Exploratory Pressure (cm H2O) at 3
                 med_ph.36 = "Medication for Pulmonary Hypertension at 36 weeks",
                 med_ph.44 = "Medication for Pulmonary Hypertension at 44 weeks")) %>%
  add_overall() %>%
  modify_header(label ~ " ") %>%
  modify_caption("**Table 3. Summary of Respiratory Support Variables**") %>%
  #modify_spanning_header(c("stat_1", "stat_2") ~ "**Gender**") %>%
  bold_labels()
```

```
# overall missingness
vis_miss(trach_df)

# missing variable summary for more than 40% missing
missingness <- trach_df %>%
  miss_var_summary() %>%
  filter(pct_miss >= 40)

#output summary table for variables missing more than 40%
kable(missingness, caption = "Variables with Significant Missingness",
      align = "l",digits = 3)


# missing variable summary grouped by medical center
#miss <- trach_df %>%
#  group_by(center) %>%
#  miss_var_summary() %>%
#  filter(pct_miss > 0 & n_miss > 20) %>%
#  arrange(pct_miss)

#output summary table for missing variable summary grouped by medical center
#kable(miss, caption = "Missingness by Medical Center",
#      align = "l",digits = 3)

# missing variable summary for medical center 2
miss2 <- trach_df %>%
  group_by(center) %>%
  miss_var_summary() %>%
  filter(center == 2 & pct_miss > 0 & n_miss > 20) %>%
  arrange(center)


#output summary table for missing variable summary for medical center 2
kable(miss2, caption = "Medical Center 2 Missing Variable Summary",
      align = "l",digits = 3)
# histograms for the independent respiratory support variables at 36 weeks
newdata <- na.omit(trach_df) %>%
  dplyr::select(weight_today.36, inspired_oxygen.36, p_delta.36, peep_cm_h2o_modified.36,
         Trach, Death) %>%
  mutate(Tracheostomy = ifelse(Trach == 1, "Yes", "No"))
  #rename("Weight at 36 Weeks" = weight_today.36,
   #      "Fraction of Inspired Oxygen at 36 weeks" = inspired_oxygen.36,
    #     "Peak Inspiratory Pressure (cmH2O) at 36 weeks" = p_delta.36,
     #    "Positive and exploratory pressure (cm H2O) at 36 weeks" = peep_cm_h2o_modified.
      #   "Tracheostomy" = Trach)
par(mar = c(1, 1, 1, 1))
par(mfrow = c(2,2))


for( i in 1:4){
  hist(newdata[,i], main = colnames(newdata)[i], xlab = colnames(newdata)[i], col = 'turqu
  mtext("Figure 1. Histograms for Continuous Independent Respiratory Support Variables
        at 36 Weeks", side = 3, line =  -2.2, outer = TRUE)
}


# histograms for the independent respiratory support variables at 44 weeks
newdata2 <- na.omit(trach_df) %>%
dplyr::select(weight_today.44, inspired_oxygen.44, p_delta.44, peep_cm_h2o_modified.44,
         Trach, Death) %>%
  mutate(Tracheostomy = ifelse(Trach == 1, "Yes", "No"))
  #rename("Weight at 44 Weeks" = weight_today.44,
   #      "Fraction of Inspired Oxygen at 44 Weeks" = inspired_oxygen.44,
    #     "Peak Inspiratory Pressure (cmH2O) at 44 Weeks" = p_delta.44,
     #    "Positive and exploratory pressure (cm H2O) at 44 Weeks" = peep_cm_h2o_modified.
      #   "Tracheostomy" = Trach)
par(mar = c(1, 1, 1, 1))
par(mfrow = c(2,2))
for( i in 1:4){
  hist(newdata2[,i], main = colnames(newdata2)[i], xlab = colnames(newdata2)[i],col = 'tur
   mtext("Figure 2. Histograms for Continuous Independent Respiratory Support Variables
        at 44 Weeks", side = 3, line =  -2.2, outer = TRUE)
}

# histograms birth weight, birth length, birth head circumference, and gestational age
newdata3 <- na.omit(trach_df) %>%
  dplyr::select(bw, birth_hc, blength, ga, hosp_dc_ga, Trach, Death) %>%
  mutate(Tracheostomy = ifelse(Trach == 1, "Yes", "No"))
  #rename("Birth Weight (g)" = bw, "Birth Head Circumference (cm)" = birth_hc,
   #      "Birth Length (cm)" = blength, "Gestational Age" = ga,
    #     "Hospital Discharge Gestational Age" = hosp_dc_ga, "Tracheostomy" = Trach)
```

```r
par(mar = c(1, 1, 1, 1))
par(mfrow = c(3,2))
for( i in 1:5){
  hist(newdata3[,i], main = colnames(newdata3)[i],xlab =colnames(newdata3)[i], col = 'turq
   mtext("Figure 3. Histograms for Continuous Independent Birth Variables", side = 3, line
}
box_plot <- function(data, x, y, title) {
  ggplot()+
  geom_boxplot(data, mapping = aes({{x}}, {{y}}, fill = {{x}}),color = "black" , outlier.c
  #geom_boxplot(data, mapping = aes({{z}}, color = {{z}}), outlier.color = "red") +
    scale_fill_viridis_d() +
    #scale_color_brewer(palette = "Dark2") +
    theme(axis.line = element_line(colour = "black",linewidth=1),
    text = element_text(size=10),
    axis.text = element_text(colour = "black",size = 10,face="bold"),
    axis.title = element_text(size = 10,face="bold"),
    axis.ticks.length=unit(.20, "cm")) +
    #axis.ticks = element_line(colour = "black", linewidth = 1))+
    theme(legend.position = "none") +
    ggtitle(title)
}

# Boxplots for Continuous Independent Respiratory Support Variables at 36 Weeks
bw36_trach <- box_plot(newdata, Tracheostomy, weight_today.36, "Weight at 36 Weeks")
inspoxy36_trach <- box_plot(newdata, Tracheostomy, inspired_oxygen.36, "Fraction of Inspir
peak36_trach <- box_plot(newdata, Tracheostomy, p_delta.36, "Peak Inspiratory Pressure (cm
posexp36_trach <- box_plot(newdata, Tracheostomy, peep_cm_h2o_modified.36, "Positive and e

plot1 <- ggarrange(bw36_trach, inspoxy36_trach, peak36_trach, posexp36_trach,
                   ncol = 4, nrow = 1)
annotate_figure(plot1, top = text_grob("Boxplots for Continuous Independent Respiratory Su

# Boxplots for Continuous Independent Respiratory Support Variables at 44 Weeks
bw44_trach <- box_plot(newdata2, Tracheostomy, weight_today.44, "Weight at 44 Weeks")
inspoxy44_trach <- box_plot(newdata2, Tracheostomy, inspired_oxygen.44, "Fraction of Inspi
peak44_trach <- box_plot(newdata2, Tracheostomy, p_delta.44, "Peak Inspiratory Pressure (c
posexp44_trach <- box_plot(newdata2, Tracheostomy, peep_cm_h2o_modified.44, "Positive and
plot2 <- ggarrange(bw44_trach, inspoxy44_trach, peak44_trach, posexp44_trach,
                   ncol = 4, nrow = 1)
annotate_figure(plot2, top = text_grob("Boxplots for Continuous Independent Respiratory Su
# Boxplots for Continuous Independent Birth Variables
aa <- box_plot(newdata3, Tracheostomy, bw, "Distribution for Birth Weight")
bb <- box_plot(newdata3, Tracheostomy, birth_hc, "Distribution for Birth Head Circumferenc
cc <- box_plot(newdata3, Tracheostomy, blength, "Distribution for Birth Lenghth")
dd <- box_plot(newdata3, Tracheostomy, ga, "Distribution for Gestational Age")
ee <- box_plot(newdata3, Tracheostomy, hosp_dc_ga, "Distribution for Hospital Discharge Ge
plot3 <- ggarrange(aa, bb, cc, dd, ee, ncol = 3, nrow = 2)
annotate_figure(plot3, top = text_grob("Boxplots for Continuous Independent Birth Variable

# subset the data for imputation
trach_subset <- trach_df %>%
  dplyr::select(c(center, bw,ga,blength,birth_hc,del_method,prenat_ster,com_prenat_ster,
       mat_chorio,gender,sga,any_surf,weight_today.36,ventilation_support_level.36,
       inspired_oxygen.36,p_delta.36,peep_cm_h2o_modified.36, med_ph.36,
       weight_today.44,ventilation_support_level_modified.44,inspired_oxygen.44,
       p_delta.44,peep_cm_h2o_modified.44,med_ph.44,hosp_dc_ga,Trach))

# impute the data set using the mice package
#apply(trach_subset, 2, function(x){return(sum(!is.na(x))/length(x))})
trach_df_mice_out <- mice(trach_subset, m=5, pri = FALSE, seed=10)

# Store each imputed data set
trach_impdf <- vector("list",5)
for (i in 1:5){
trach_impdf[[i]] <- mice::complete(trach_df_mice_out,i)
}

logistic <- function(df) {
  #' Runs 10-fold CV for lasso and returns corresponding coefficients
  #' @param df, data set
  #' @return coef, coefficients for minimum cv error

  # Matrix form for ordered variables
  x.ord <- model.matrix(Trach ~., data = df)[,-1]
  y.ord <- df$Trach

  # Logistic model
  logistic_mod <- glm(y.ord ~ x.ord, family = "binomial")

  # Get coefficients
  coef <- coef(logistic_mod)
  return(coef)
}

# Find average lasso coefficients over imputed datasets
logistic_coef1 <- logistic(trach_impdf[[1]])
logistic_coef2 <- logistic(trach_impdf[[2]])
logistic_coef3 <- logistic(trach_impdf[[3]])
logistic_coef4 <- logistic(trach_impdf[[4]])
logistic_coef5 <- logistic(trach_impdf[[5]])
logistic_coef <- cbind(logistic_coef1, logistic_coef2, logistic_coef3,
logistic_coef4, logistic_coef5)
avg_coefs_logistic <- apply(logistic_coef, 1, mean)
names(avg_coefs_logistic) <- gsub(pattern = "x.ord", replacement = "", x = names(avg_coefs

# Find predicted probabilities on long imputed data (no rounding applied in this case!)
trach_df_long <- mice::complete(trach_df_mice_out,action="long")
x_vars <- model.matrix(Trach~ ., trach_df_long)[,-c(2,3)]
trach_df_long$logistic_score <- x_vars %*% avg_coefs_logistic
mod_logistic <- glm(Trach~logistic_score, data = trach_df_long, family = "binomial")
predict_probs_logistic <- predict(mod_logistic, type="response")

lasso <- function(df) {
  #' Runs 10-fold CV for lasso and returns corresponding coefficients
  #' @param df, data set
  #' @return coef, coefficients for minimum cv error

  # Matrix form for ordered variables
  x.ord <- model.matrix(Trach~., data = df)[,-1]
  y.ord <- df$Trach

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

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

```r
  coef <- coef(lasso_mod)
  return(coef)
}


# Find average lasso coefficients over imputed datasets
lasso_coef1 <- lasso(trach_impdf[[1]])
lasso_coef2 <- lasso(trach_impdf[[2]])
lasso_coef3 <- lasso(trach_impdf[[3]])
lasso_coef4 <- lasso(trach_impdf[[4]])
lasso_coef5 <- lasso(trach_impdf[[5]])
lasso_coef <- cbind(lasso_coef1, lasso_coef2, lasso_coef3,
lasso_coef4, lasso_coef5)
avg_coefs_lasso <- apply(lasso_coef, 1, mean)

# Find predicted probabilities on long imputed data (no rounding applied in this case!)
trach_df_long <- mice::complete(trach_df_mice_out,action="long")
x_vars <- model.matrix(Trach~. , trach_df_long)[,-c(2,3)]
trach_df_long$lasso_score <- x_vars %*% avg_coefs_lasso
mod_lasso <- glm(Trach~lasso_score, data = trach_df_long, family = "binomial")
predict_probs_lasso <- predict(mod_lasso, type="response")
ridge <- function(df) {
#' Runs 10-fold CV for lasso and returns corresponding coefficients
#' @param df, data set
#' @return coef, coefficients for minimum cv error

# Matrix form for ordered variables
x.ord <- model.matrix(Trach~., data = df)[,-1]
y.ord <- df$Trach

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

# Ridge model
ridge_mod <- cv.glmnet(x.ord, y.ord, nfolds = 10, foldid = folds,
alpha = 0, family = "binomial")

# Get coefficients
coef <- coef(ridge_mod, lambda = ridge_mod$lambda.min)
```

```
return(coef)
}
# Find average lasso coefficients over imputed datasets
ridge_coef1 <- ridge(trach_impdf[[1]])
ridge_coef2 <- ridge(trach_impdf[[2]])
ridge_coef3 <- ridge(trach_impdf[[3]])
ridge_coef4 <- ridge(trach_impdf[[4]])
ridge_coef5 <- ridge(trach_impdf[[5]])
ridge_coef <- cbind(ridge_coef1, ridge_coef2, ridge_coef3,
ridge_coef4, ridge_coef5)
avg_coefs_ridge <- apply(ridge_coef, 1, mean)
#avg_coefs_ridge <- avg_coefs_ridge[names(avg_coefs_ridge) != "(Intercept)" & avg_coefs_ri
#med_avg_coefs_ridge <- round(avg_coefs_ridge/median(avg_coefs_ridge))

# Find predicted probabilities on long imputed data (no rounding applied in this case!)
trach_df_long <- mice::complete(trach_df_mice_out,action="long")
x_vars <- model.matrix(Trach ~ ., trach_df_long)[,-c(2,3)]
trach_df_long$ridge_score <- x_vars %*% avg_coefs_ridge
mod_ridge <- glm(Trach~ ridge_score, data = trach_df_long, family = "binomial")
predict_probs_ridge <- predict(mod_ridge, type="response")

ggplot() +
  geom_histogram(aes(x=predict_probs_ridge, fill=as.factor(mod_ridge$y)),
                bins=30) +
  scale_fill_discrete(name="Tracheostomy") +
  labs(x="Predicted Probabilities", y="Count")
# data frame for regression coefficients
coeff_df <- data.frame(coefs_logistic = avg_coefs_logistic,
                       coefs_lasso =avg_coefs_lasso,
                       coefs_ridge = avg_coefs_ridge)

# output table for regression coefficients
kable(coeff_df, caption = "Regression Coefficient Comparisons", align = "l",digits = 3)
# performance evaluation using discrimination, ROC Curve
par(mfrow=c(1,3))

roc_mod_logistic <- roc(predictor=predict_probs_logistic, type="response",
                    response=as.factor(trach_df_long$Trach),
                    levels = c(0,1), direction = "<")
plot(roc_mod_logistic, print.auc=TRUE, print.thres = TRUE, main=list("Logistic ROC Curve")
roc_mod_lasso <- roc(predictor=predict_probs_lasso,
response=as.factor(trach_df_long$Trach),
levels = c(0,1), direction = "<")
plot(roc_mod_lasso, print.auc=TRUE, print.thres = TRUE, main=list("Lasso ROC Curve"))

roc_mod_ridge <- roc(predictor=predict_probs_ridge,
response=as.factor(trach_df_long$Trach),
levels = c(0,1), direction = "<")
plot(roc_mod_ridge, print.auc=TRUE, print.thres = TRUE, main=list("Ridge ROC Curve"))

#
pred_ys_logistic <- ifelse(predict_probs_logistic > 0.166, 1, 0)
tab_outcome_logistic <- table(trach_df_long$Trach, pred_ys_logistic)
pred_ys_lasso <- ifelse(predict_probs_lasso > 0.190, 1, 0)
tab_outcome_lasso <- table(trach_df_long$Trach, pred_ys_lasso)
pred_ys_ridge <- ifelse(predict_probs_ridge > 0.132, 1, 0)
tab_outcome_ridge <- table(trach_df_long$Trach, pred_ys_ridge)

# logistic performance evaluation measures
sens_log <- tab_outcome_logistic[2,2]/(tab_outcome_logistic[2,1]+tab_outcome_logistic[2,2]
spec_log <- tab_outcome_logistic[1,1]/(tab_outcome_logistic[1,1]+tab_outcome_logistic[1,2]
ppv_log <- tab_outcome_logistic[2,2]/(tab_outcome_logistic[1,2]+tab_outcome_logistic[2,2])
npv_log <- tab_outcome_logistic[1,1]/(tab_outcome_logistic[1,1]+tab_outcome_logistic[2,1])
acc_log <- (tab_outcome_logistic[1,1]+tab_outcome_logistic[2,2])/sum(tab_outcome_logistic)
#rbind(round(c(sens_log, spec_log, ppv_log, npv_log, acc_log), 3))

# lasso performance evaluation measures
sens_lasso <- tab_outcome_lasso[2,2]/(tab_outcome_lasso[2,1]+tab_outcome_lasso[2,2])
spec_lasso <- tab_outcome_lasso[1,1]/(tab_outcome_lasso[1,1]+tab_outcome_lasso[1,2])
ppv_lasso <- tab_outcome_lasso[2,2]/(tab_outcome_lasso[1,2]+tab_outcome_lasso[2,2])
npv_lasso <- tab_outcome_lasso[1,1]/(tab_outcome_lasso[1,1]+tab_outcome_lasso[2,1])
acc_lasso <- (tab_outcome_lasso[1,1]+tab_outcome_lasso[2,2])/sum(tab_outcome_lasso)
#rbind(round(c(sens_lasso, spec_lasso, ppv_lasso, npv_lasso, acc_lasso), 3))

# ridge performance evaluation measures
sens_ridge <- tab_outcome_ridge[2,2]/(tab_outcome_ridge[2,1]+tab_outcome_ridge[2,2])
spec_ridge <- tab_outcome_ridge[1,1]/(tab_outcome_ridge[1,1]+tab_outcome_ridge[1,2])
ppv_ridge <- tab_outcome_ridge[2,2]/(tab_outcome_ridge[1,2]+tab_outcome_ridge[2,2])
npv_ridge <- tab_outcome_ridge[1,1]/(tab_outcome_ridge[1,1]+tab_outcome_ridge[2,1])
acc_ridge <- (tab_outcome_ridge[1,1]+tab_outcome_ridge[2,2])/sum(tab_outcome_ridge)
#rbind(round(c(sens_ridge, spec_ridge, ppv_ridge, npv_ridge, acc_ridge), 3))
```

```r
# data frame for model performance measures
measures <- data.frame(Model = c("Logistic", "Lasso", "Ridge"),
  Sensitivity = (round(c(sens_log, sens_lasso, sens_ridge), 3)),
  Specificity = (round(c(spec_log, spec_lasso, spec_ridge), 3)),
  PPV = (round(c(ppv_log, ppv_lasso, ppv_ridge), 3)),
  NPV = (round(c(npv_log, npv_lasso, npv_ridge), 3)),
  Accuracy = (round(c(acc_log, acc_lasso, acc_ridge), 3)))

# output table for model performance measures
#measures %>%
#mutate_all(linebreak) %>%
#kbl(caption = "Model Performance Measures",
#col.names=linebreak(c("Model", "Sensitivity", "Specificity", "Positive Predictive Values"
#                      "Negative Predictive Values", "Overall Accuracy")),
#booktabs=T, escape=F, align = "c") %>%
#kable_styling(full_width = FALSE, latex_options = c('hold_position'))
kable(measures, caption = "Model Performance Measures", align = "l",digits = 3)

num_cuts <- 10

# logistic calibration metrics
log_calib_data <- data.frame(prob = predict_probs_logistic,
                         bin = cut(predict_probs_logistic, breaks = num_cuts),
                         class = trach_df_long$Trach)
log_calib_data <- log_calib_data %>%
  group_by(bin) %>%
  summarize(observed = sum(as.numeric(as.character(class)))/n(),
            expected = sum(prob)/n(),
            se = sqrt(observed*(1-observed)/n()))

# calibration plot for logistic
log_calib_plot <- ggplot(log_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", title = "Logistic") +
  theme_minimal()

# lasso calibration metrics
```

```r
lasso_calib_data <- data.frame(prob = predict_probs_lasso,
                               bin = cut(predict_probs_lasso, breaks = num_cuts),
                               class = trach_df_long$Trach)
lasso_calib_data <- lasso_calib_data %>%
  group_by(bin) %>%
  summarize(observed = sum(as.numeric(as.character(class)))/n(),
            expected = sum(prob)/n(),
            se = sqrt(observed*(1-observed)/n()))

# calibration plot for lasso
lasso_calib_plot <- ggplot(lasso_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", title = "Lasso") +
  theme_minimal()

# ridge calibration metrics
ridge_calib_data <- data.frame(prob = predict_probs_ridge,
                               bin = cut(predict_probs_ridge, breaks = num_cuts),
                               class = trach_df_long$Trach)
ridge_calib_data <- ridge_calib_data %>%
  group_by(bin) %>%
  summarize(observed = sum(as.numeric(as.character(class)))/n(),
            expected = sum(prob)/n(),
            se = sqrt(observed*(1-observed)/n()))

# calibration plot for ridge
ridge_calib_plot <- ggplot(ridge_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", title = "Ridge") +
  theme_minimal()

grid.arrange(log_calib_plot, lasso_calib_plot, ridge_calib_plot, nrow = 1, ncol = 3)
# performance evaluation using Brier Score
```

```
b1 <- BrierScore(as.numeric(as.character(trach_df_long$Trach)), predict_probs_logistic)
b2 <- BrierScore(as.numeric(as.character(trach_df_long$Trach)), predict_probs_lasso)
b3 <- BrierScore(as.numeric(as.character(trach_df_long$Trach)), predict_probs_ridge)
brier <- data.frame(Models = c("Logistic", "Lasso", "Ridge"),
  BrierScore = (round(c(b1, b2, b3), 3)))

#output table performance evaluation using Brier Score
#brier %>%
#mutate_all(linebreak) %>%
#kbl(caption = "Brier Score Measures",
#col.names=linebreak(c("Model", "Brier Score")),
#booktabs=T, escape=F, align = "c") %>%
#kable_styling(full_width = FALSE, latex_options = c('hold_position'))
kable(brier, caption = "Brier Score Measures", align = "l",digits = 3)
```