PHP 2550 Project 2

Victoria Grase

2023-11-12

Abstract

This study aims to develop predictive models for the composite outcome of tracheostomy, with a focus on guiding indication criteria and timing of tracheostomy placement. The analysis incorporates Lasso regression, logistic regression, backward stepwise regression, and multiple imputation, to enhance the accuracy and generalizability of the models. The complete dataset comprises diverse variables such as patient demographics, clinical indicators, and relevant medical history. The Lasso regression, known for its feature selection capability, is employed to identify the most influential predictors associated with tracheostomy. Logistic regression further refines the model by capturing the probability of tracheostomy occurrence based on the identified predictors. Backward stepwise regression aids in systematically refining the model by iteratively eliminating less impactful variables. Recognizing the challenges posed by missing data, multiple imputation techniques are implemented to handle incomplete information effectively. This involves creating imputed datasets, performing analyses on each, and viewing results to provide and compare between regressions a comprehensive and robust prediction model. The developed models aim to contribute valuable insights into the factors influencing the composite outcome of tracheostomy, providing a foundation for evidence-based decision-making in clinical settings. The combination of Lasso, logistic regression, backward stepwise regression, and multiple imputation techniques ensures a thorough analysis of the dataset, offering a powerful tool for clinicians and researchers alike in understanding and predicting tracheostomy outcomes.

Introduction

The primary objective of this investigation is to construct a robust regression model tailored to predict the composite outcome of tracheostomy, thereby offering valuable guidance on the indication criteria and optimal timing for tracheostomy placement. A distinctive aspect of this model development lies in its focus on statistical variables collected at both 36 and 44 weeks of gestation, providing a comprehensive understanding of factors influencing the likelihood of tracheostomy. The variables of particular interest encompass birth-related factors such as weight, gestational age, the administration of prenatal steroids for lung development, maternal race, gender (acknowledging potential gender-based differences in outcomes), and the presence of chorioamnionitis, an infection of the amniotic fluid. Additionally, variables recorded at 36 and 44 weeks, including weight and the eventual need for tracheostomy, contribute crucial insights into the predictive dynamics.

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

additional risk factors important for the tracheostomy outcome prediction.

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

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

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

```
##     record_id center mat_race mat_ethn  bw ga blength birth_hc del_method
## 789   2000824      2        0        2 780 28      34       24          2
## 790   2000824      2        0        2 780 28      34       24          2
## 791   2000824      2        0        2 780 28      34       24          2
## 792   2000824      2        0        2 780 28      34       24          2
##     prenat_ster com_prenat_ster mat_chorio gender     sga any_surf
## 789         Yes             Yes         No   Male Not SGA      Yes
## 790         Yes             Yes         No   Male Not SGA      Yes
## 791         Yes             Yes         No   Male Not SGA      Yes
## 792         Yes             Yes         No   Male Not SGA      Yes
##     weight_today.36 ventilation_support_level.36 inspired_oxygen.36 p_delta.36
## 789            1825                            1               0.26          0
## 790            1825                            1               0.26          0
## 791            1825                            1               0.26          0
## 792            1825                            1               0.26          0
##     peep_cm_h2o_modified.36 med_ph.36 weight_today.44
## 789                       7         0              NA
```

```
## 790                       7         0             NA
## 791                       7         0             NA
## 792                       7         0             NA
##     ventilation_support_level_modified.44 inspired_oxygen.44 p_delta.44
## 789                                    NA                NA         NA
## 790                                    NA                NA         NA
## 791                                    NA                NA         NA
## 792                                    NA                NA         NA
##     peep_cm_h2o_modified.44 med_ph.44 hosp_dc_ga Trach Death
## 789                      NA        NA       38.7     0    No
## 790                      NA        NA       38.7     0    No
## 791                      NA        NA       38.7     0    No
## 792                      NA        NA       38.7     0    No
```

```
## # A tibble: 30 x 3
##    variable                               n_miss pct_miss
##    <chr>                                   <int>    <dbl>
##  1 inspired_oxygen.44                        448     45.0
##  2 p_delta.44                                448     45.0
##  3 weight_today.44                           446     44.8
##  4 peep_cm_h2o_modified.44                   446     44.8
##  5 any_surf                                  433     43.5
##  6 ventilation_support_level_modified.44     424     42.6
##  7 med_ph.44                                 424     42.6
##  8 com_prenat_ster                           193     19.4
##  9 p_delta.36                                128     12.9
## 10 hosp_dc_ga                                124     12.4
## # i 20 more rows
```

```
##     record_id   pct_na
## 1     1000003 36.66667
## 13    1000027 40.00000
## 15    1000034 40.00000
## 26    1000049 43.33333
## 37    1000062 36.66667
## 57    1000250 40.00000
## 73   12-000043 36.66667
## 81   12-000062 36.66667
## 86   12-000076 36.66667
## 95   12-000116 36.66667
## 100 12-000133 40.00000
## 126 12-000204 36.66667
## 244   2000081 46.66667
## 284   2000133 40.00000
## 412   2000304 40.00000
## 435   2000329 36.66667
## 450   2000344 36.66667
## 657   2000586 43.33333
## 876   4000010 36.66667
## 904   4000047 36.66667
## 908   4000051 36.66667
## 919   4000063 36.66667
## 921   4000065 40.00000
## 926   4000071 36.66667
## 973   7000011 36.66667
```

Table 1: Summary of Demographics and Infant Delivery

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

```
## 978   7000029 36.66667
```

Study Characteristics and Population

The dataset comprises information from 996 participants encompassing a total of 28 variables, offering a comprehensive representation of demographics, gestational age, birthweight, prenatal steroid administration (facilitating lung development), maternal race/ethnicity, gender (acknowledging gender-based differences), and the presence of chorioamnionitis (an infection of the amniotic fluid). Additionally, the study captures measurements at both 36 and 44 weeks for critical variables such as Peak Inspiratory Pressure, Positive End Exploratory Pressure, Medication for Pulmonary Hypertension, weight, ventilation support level, and fraction of inspired oxygen. Within this wealth of data, binary potential outcomes were recorded, presenting opportunities for evaluating the goal of determining optimal tracheostomy placement timing, with death and the presence of tracheostomy being key outcomes of interest. To streamline the analysis and enhance interpretability, the focus was deliberately narrowed down to the binomial outcome variable "Trach" to discern the best-fitting model for achieving the study's goals.

Table 3 contains variables related to oxygen levels, pressure, and gestational age, comparing the overall cohort (N = 996) with subgroups based on tracheostomy placement (Trach 0, N = 850; Trach 1, N = 146).

Table 2: Summary of Infant Information

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

Table 3: ** Average Birth and Infants Weights and Heights**

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

Table 4: **Average of Oxygen, Pressure, and Gestiational Age**

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

Table 5: Maternal Demographics

	Overall, N = 939	**Hispanic or Latino**, N = 74	**Not Hispanic or Latino
___Race of Mother___			
American Indian or Alaskan Native	534 (57%)	22 (32%)	512 (59%)
Asian	286 (31%)	2 (2.9%)	284 (33%)
Other	111 (12%)	44 (65%)	67 (7.8%)
Missing	8	6	2
___severity___			
1	236 (27%)	21 (32%)	215 (27%)
2	228 (26%)	12 (18%)	216 (27%)
3	399 (46%)	32 (49%)	367 (46%)
Missing	76	9	67

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

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

The Fraction of Inspired Oxygen (FiO2) at 36 weeks demonstrated a mean of 0.34 in the overall cohort, with a slight decrease in the tracheostomy group (0.32) and a notable increase in the FiO2 for infants with tracheostomy (0.49). Similarly, at 44 weeks, the FiO2 means remained consistent, but the tracheostomy group exhibited a slightly lower mean (0.44). For Peak Inspiratory Pressure at 36 and 44 weeks, the overall averages increased from 5 to 8 cmH2O, and 15 to 21 cmH2O, respectively. The tracheostomy group consistently demonstrated higher pressures at both time points compared to the non-tracheostomy group. Hospital Discharge Gestational Age showcased a mean of 53 weeks for the overall cohort, while the tracheostomy and non-tracheostomy groups had mean values of 49 and 80 weeks, respectively.

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

Table 5 provides a comprehensive overview of key factors related to tracheostomy placement in a cohort of 996 infants. Among them, 850 infants did not require a tracheostomy, while 146 infants did. The gender distribution showed a predominance of males in both groups. Prenatal steroid administration was prevalent, with 87% of infants receiving steroids in the non-tracheostomy group and 94% in the tracheostomy group. Delivery methods varied, with 71% of infants in the non-tracheostomy group delivered via c-section, compared to 79% in the tracheostomy group. Ventilation support levels at 36 weeks revealed a higher proportion of infants requiring support in the tracheostomy group. This trend continued at 44 weeks, indicating a significant association between tracheostomy placement and the need for prolonged ventilation. Additionally, medication for pulmonary hypertension at 36 and 44 weeks had a higher prevalence in the tracheostomy group. The severity of (BPD) also exhibited distinct patterns, with a higher proportion of severe cases in the tracheostomy group.

Methods

Initially, a subset of variables deemed critical for the analysis is selected and appropriately renaTo address missing values, the Multiple Imputation by Chained Equations (MICE) method is employed, generating five imputed datasets to ensure robustness and account for uncertainty in the imputation process. A random seed is set to enhance reproducibility. The imputed datasets are stored in a list for further analysis, and the mice::complete function is utilized to obtain the completed datasets for each imputation iteration. The entire imputation process is then saved as an RDS file to preserve the imputed datasets for future analyses.

Visualization of the imputed datasets is shown through a plot looking at distribution and patterns of imputed values. The imputation methods used can be inspected through the trache_mice$method attribute. Finally, the imputed datasets are made into a dataframe named completed_data using the complete function, which aggregates the imputed values into a single dataset, ready for subsequent analyses. This imputation process is essential for ensuring a more robust and complete dataset, enhancing the reliability of statistical analyses and model building in which we will need for the following regression: lasso, logistic, forward, and backward.

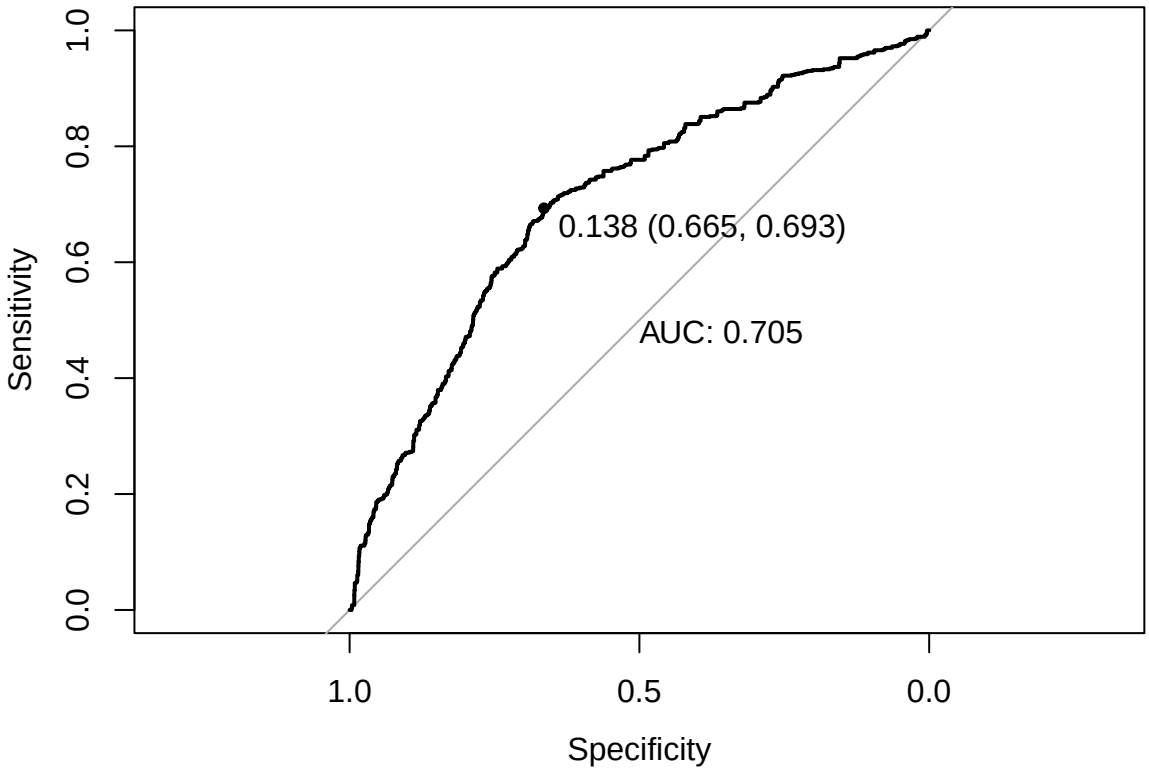
Model Descriptions

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

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

The logistic_backward function is designed to perform backward stepwise logistic regression on a given dataset. This process involves fitting a logistic regression model with all potential predictor variables and then iteratively removing variables that do not contribute significantly to the model.Firstly, the function removes irrelevant variables then, it creates a matrix of predictor variables and the response variable needed for logistic regression. Two models are fitted: one which represents the full logistic regression model with all potential predictors, and another, which represents the null model with only the intercept. The function then utilizes the step function to perform backward stepwise variable selection. It specifies the direction as 'backward' and uses the AIC criterion to determine the best-fitting model. The results of the variable selection process are stored in the backwardstep object. Finally, the function extracts the coefficients from the selected model (backwardstep) and returns them along with the model object. This allows users to access both the coefficients and the model for further analysis or interpretation.

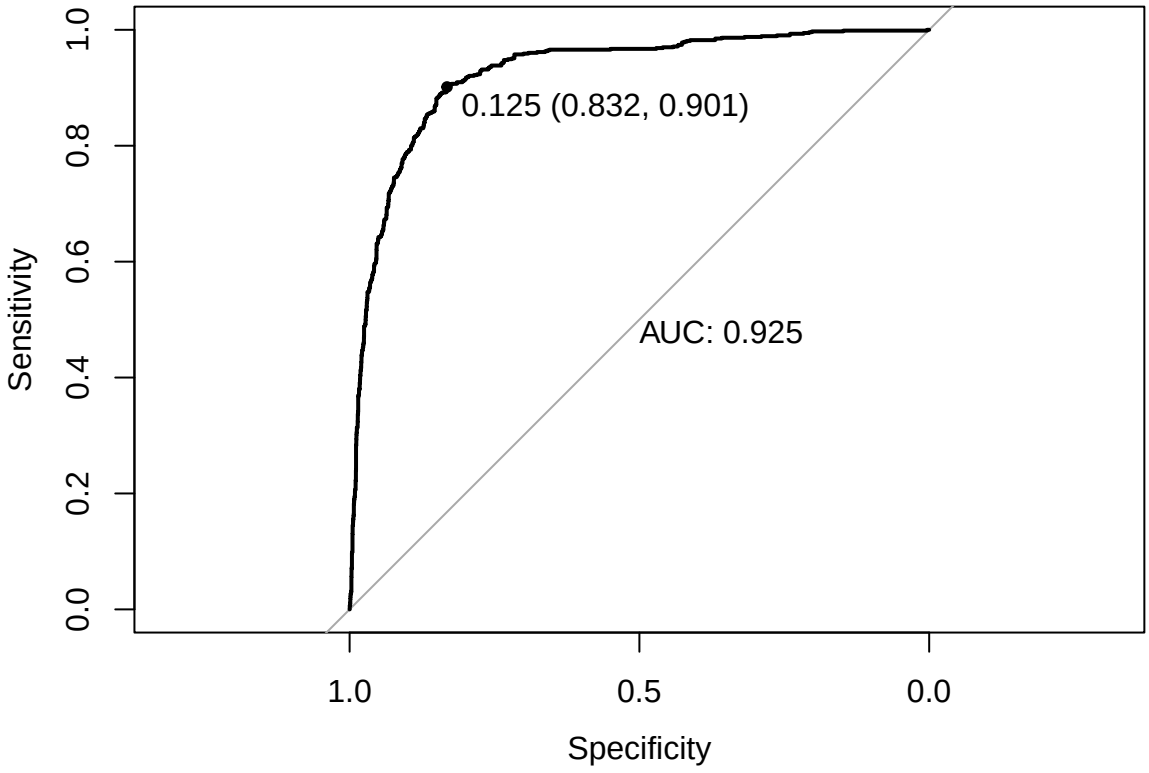
Statistical Analysis




```
##    threshold  specificity sensitivity
## 1       -Inf 0.0000000000           1
## 2 0.05066299 0.0002352941           1
## 3 0.05067792 0.0004705882           1
## 4 0.05096021 0.0007058824           1
## 5 0.05125170 0.0009411765           1
## 6 0.05131917 0.0011764706           1
```

```
##      threshold specificity sensitivity
## 1664 0.1174883   0.5618824   0.7506849
```

```
##    pred_ys
##        0    1
##   0 2378 1872
##   1  177  553
```




```
##      threshold  specificity sensitivity
## 1         -Inf 0.0000000000           1
## 2 5.838621e-07 0.0002352941           1
## 3 1.108565e-06 0.0011764706           1
## 4 1.534638e-06 0.0014117647           1
## 5 1.557653e-06 0.0016470588           1
## 6 1.595642e-06 0.0018823529           1
```

```
##      threshold specificity sensitivity
## 2815 0.2885031   0.9167059   0.7506849
```

```
##    pred_ys_log
##        0    1
##   0 3896  354
##   1  182  548
```

In summary, the Logistic Regression model with an AUC of 0.925 is considered better than the Lasso model with an AUC of 0.705. A higher AUC indicates that the Logistic Regression model has a better ability to distinguish between positive and negative instances.In addition a higher AUC value is generally associated with improved predictive performance. It suggests that the Logistic Regression model has a better balance between sensitivity and specificity, making it more reliable in making accurate predictions across different thresholds.The AUC summarizes the model's performance across various classification thresholds, providing a comprehensive view of its ability to trade off true positive rate and false positive rate. The higher AUC for Logistic Regression indicates a better performance across a range of decision thresholds compared to Lasso.

The Precision (PPV): The Logistic Regression model outperforms the Lasso model significantly in terms of Precision, indicating a higher proportion of true positive predictions among the positive predictions made by the model. The Negative Predictive Value (NPV): The Logistic Regression model also exhibits a higher

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

Discussion

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

Limitations

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

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

References:

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

Code Appendix

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

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

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

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

overall_missing<-miss_var_summary(project2)
print(overall_missing)

#Look at missing data for each parent
pct_na_r <- rowSums(is.na(project2)) / ncol(project2) * 100
row_na <- data.frame(record_id = project2$record_id, pct_na = pct_na_r)
row_na <- row_na[row_na$pct_na > 35,]
print(row_na)
#Create severity of BPD variable indicator at week 36
project2<-project2%>% mutate(severity = case_when(
  (ventilation_support_level.36 == 0 |
     (ventilation_support_level.36 == 1  & inspired_oxygen.36 < 0.22)) ~ 'Mild',
  ((ventilation_support_level.36 == 2 & inspired_oxygen.36 <= 0.21) |
     (ventilation_support_level.36 == 1 & inspired_oxygen.36 < 0.30 & inspired_oxygen.36 >= 0.22)) ~ 'M
   ((ventilation_support_level.36 == 2 & inspired_oxygen.36 > 0.21) |
```

```
      (ventilation_support_level.36 == 1 & inspired_oxygen.36 >= 0.30)) ~ 'Severe'))
# which(is.na(project2$severity))
# sum(is.na(project2$severity))

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


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

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

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

tbl_butchered2<-table2%>%
  tbl_butcher()
tbl_butchered2

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

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

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

tbl_butchered3<-table3%>%
  tbl_butcher()
tbl_butchered3

# average tracheostomy at discharge and death before discharge
table4<-project2 %>%
  select(Trach, gender, prenat_ster,del_method,ventilation_support_level.36,ventilation_support_level_m
  tbl_summary(
```

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

tbl_butchered4<-table4%>%
  tbl_butcher()
tbl_butchered4
# Select variables and rename
# project2<- project2 %>% select(c(center,bw,ga,blength,birth_hc,del_method,prenat_ster,com_prenat_ster
#Create imputation for variables at 44 weeks including weight, Fraction of Inspired Oxygen,Peak Inspira

#Perform multiple imputation (generating 5 imputed datasets)
# Set a random seed for reproducibility
project2<- project2[,-c(31,32,33)]
trache_mice<- mice(project2, m = 5, print = FALSE, seed = 10)
# Store each imputed data set
trache_imputed<- list()
for (i in 1:5){
  trache_imputed[[i]]<-complete(trache_mice,i)
}
#saveRDS(trache_mice,file="trache_mice_final_1.RDS")
#trache_mice<- readRDS("trache_mice_final_1.RDS")
#plot(trache_mice)
#trache_mice$method
#View the imputed datasets in the
completed_data<-complete(trache_mice)
#columns_to_remove <-c(30,31,32)
#completed_data<-completed_data[,-columns_to_remove ]

#overall_missing1<-miss_var_summary(completed_data)
#overall_missing1

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

```
   #grid range for lambda
  grid<- 10^ seq (10 , -2, length = 100)
  df<- df[,c(-1,-30)] #remove death

  # Matrix form for ordered variables
  x.ord.2<- model.matrix(Trach~.^2, data =df)[,-c(29)]
  x.ord <- model.matrix(Trach~., data = df)[,-c(29)]
  y.ord <- df$Trach
#   # Assuming you have a data frame or matrix named 'data'
# # Subsample 'y' to match the number of rows in 'x'
#   #set.seed(123)  # for reproducibility
#
  # Generate folds
  k <- 10
  set.seed(1) # consistent seeds between imputed data sets
  folds <- sample(1:k, nrow(df), replace=TRUE)

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

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

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

#
# Find average lasso coefficients ovetr imputed datasets
lasso_coef1 <- lasso(trache_imputed[[1]])
lasso_coef2 <- lasso(trache_imputed[[2]])
lasso_coef3 <- lasso(trache_imputed[[3]])
lasso_coef4 <- lasso(trache_imputed[[4]])
lasso_coef5 <- lasso(trache_imputed[[5]])

lasso_coef_all1<- cbind(lasso_coef1[[1]], lasso_coef2[[1]], lasso_coef3[[1]],lasso_coef4[[1]], lasso_co
lasso_coef_all2<- cbind(lasso_coef1[[2]], lasso_coef2[[2]], lasso_coef3[[2]],lasso_coef4[[2]], lasso_co
#
#for the model with  main interactions
avg_coefs_lasso1<- apply(lasso_coef_all1, 1, mean)
var_coefs_lasso1<- apply(lasso_coef_all1, 1, var)
#for the model with two-way interactions
```

```r
avg_coefs_lasso2<- apply(lasso_coef_all2, 1, mean)
var_coefs_lasso2<- apply(lasso_coef_all2, 1, var)
#
# Find predicted probabilities on long imputed data (no rounding applied in this case!)
trach_df_long<-complete(trache_mice,action="long")
subset_long<- trach_df_long[,-c(1,2,3,32)]
x_vars <- model.matrix(Trach~. , subset_long)
subset_long$lasso_score<- x_vars%*% (avg_coefs_lasso1)
mod_lasso<- glm(Trach~lasso_score, data = subset_long, family = "binomial")
predict_probs_lasso<- predict(mod_lasso, type="response")

#for two way interactions
# x_vars2<-model.matrix(Trach~(.)^2, subset_long[,-30])
# subset_long$lasso_score2<- x_vars2 %*% (avg_coefs_lasso2)
# mod_lasso2<- glm(Trach~lasso_score2, data = subset_long, family = "binomial")
# predict_probs_lasso2 <- predict(mod_lasso2, type="response")
#Binomial logistic regression
logistic <- function(df) {
  df<- df[,c(-1,-30)] #remove death

  x.ord<-model.matrix(Trach~., data = df)[,-c(1)]
  x.ord.2<-model.matrix(Trach~.^2, data =df)[,-c(1)]
  y.ord<-df$Trach

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

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

#return best model
logstic_models_coef<- list()
logistic_models_coef<- list(coef1_log, coef2_log, logistic_mod,logistic_mod1)
return(logistic_models_coef)
}

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

```r
# Find predicted probabilities on long imputed data (no rounding applied in this case!)
trach_df_long<- complete(trache_mice,action="long")
subset_long<- trach_df_long[,-c(1,2,3,32)]
x_vars_log<- model.matrix(Trach~.,subset_long)
subset_long$logistic<-x_vars_log%*% avg_coefs_log1
mod_logistic<- glm(Trach~logistic, data = subset_long, family = "binomial")
predict_probs_logistic<- predict(mod_logistic, type="response")
library(MASS)
logistic_backward <- function(df)
 {
   #' Fits logistic model and returns corresponding coefficients
   #' @param df, data set
   #' @return coef, model coefficients

  df<- df[,c(-1,-30)] #remove death
   # Matrix form for ordered variables
  #including two-way interactions
  #x.ord2 <- model.matrix(Trach~.^2, data = train)[,-c(29)] #remove trach
  #only single terms
  x.ord <- model.matrix(Trach~., data = df)[,-c(29)]
  y.ord <- df$Trach

  log_modfull <- glm(y.ord ~ x.ord, family = "binomial")
   log_modnull <- glm(y.ord ~ 1, family = "binomial")

backwardstep <- step(log_modfull, scope = formula(log_modnull), direction='backward', trace = 0)
coef <- coef(backwardstep)
#for one way interactions
predict_probs_backwardstep<- predict(backwardstep, type="response", newdata = subset_long)

dat <- list(coef, backwardstep)
return(dat)

}
library(pROC)
lasso_plot<-ggplot() +
  geom_histogram(aes(x=predict_probs_lasso, fill=as.factor(mod_lasso$y)),
                 bins=30) +
  scale_fill_discrete(name="Tracheostomy Presence") +
  labs(x="Predicted Probabilities Lasso", y="Count")
roc_mod_lasso<- roc(predictor=predict_probs_lasso,
                response=as.factor(mod_lasso$y),
                levels = c(0,1), direction = "<")
plot(roc_mod_lasso, print.auc=TRUE, print.thres = TRUE)
roc_vals_lasso<- coords(roc=roc_mod_lasso, x = "all")
head(roc_vals_lasso)
roc_vals_lasso[roc_vals_lasso$sensitivity > 0.75, ] %>% tail(n=1)

pred_ys <- ifelse(predict_probs_lasso > 0.117, 1, 0)
pred_ys <- factor(pred_ys, levels = c("0", "1"))
tab_outcome <- table(mod_lasso$y, pred_ys)
tab_outcome
#sens <- tab_outcome[2,2]/(tab_outcome[2,1]+tab_outcome[2,2])
```

```r
#spec <- tab_outcome[1,1]/(tab_outcome[1,1]+tab_outcome[1,2])
ppv <- tab_outcome[2,2]/(tab_outcome[1,2]+tab_outcome[2,2])
npv <- tab_outcome[1,1]/(tab_outcome[1,1]+tab_outcome[2,1])
acc <- (tab_outcome[1,1]+tab_outcome[2,2])/sum(tab_outcome)
vals <- data.frame(Measures = c("PPV", "NPV", "Acc"),
           Values = round(c(ppv, npv, acc),3))


log_plot<-ggplot() +
  geom_histogram(aes(x=predict_probs_logistic, fill=as.factor(mod_logistic$y)),
                 bins=30) +
  scale_fill_discrete(name="Tracheostomy Presence") +
  labs(x="Predicted Probabilities Logistic", y="Count")
roc_mod_logistic<- roc(predictor=predict_probs_logistic,
               response=as.factor(mod_logistic$y),
               levels = c(0,1), direction = "<")
plot(roc_mod_logistic, print.auc=TRUE, print.thres = TRUE)
roc_vals_logistic<- coords(roc=roc_mod_logistic, x = "all")
head(roc_vals_logistic)
roc_vals_logistic[roc_vals_logistic$sensitivity > 0.75, ] %>% tail(n=1)
pred_ys_log<- ifelse(predict_probs_logistic > 0.289, 1, 0)
pred_ys_log<- factor(pred_ys_log, levels = c("0", "1"))
tab_outcome_log<-table(mod_logistic$y, pred_ys_log)
tab_outcome_log
#sens <- tab_outcome[2,2]/(tab_outcome[2,1]+tab_outcome[2,2])
#spec <- tab_outcome[1,1]/(tab_outcome[1,1]+tab_outcome[1,2])
ppv_log<- tab_outcome_log[2,2]/(tab_outcome_log[1,2]+tab_outcome_log[2,2])
npv_log<- tab_outcome_log[1,1]/(tab_outcome_log[1,1]+tab_outcome_log[2,1])
acc_log<- (tab_outcome_log[1,1]+tab_outcome_log[2,2])/sum(tab_outcome_log)
vals_log<- data.frame(Measures = c("PPV", "NPV", "Acc"), Values = round(c(ppv_log, npv_log, acc_log),3))


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

```
# vals_back<- data.frame(Measures = c("PPV", "NPV", "Acc"), Values = round(c(ppv_back, npv_back, acc_ba
```