Project_2_PDA

Yu Yan

2023-10-31

Introduction

This report incorporates development of statistical regression models that aim to predict occurrence of tracheotomy placement or death with respect to the symptom of severe bronchopulmonary dysplasia (sBPD). The study population was drawn from BPD Collaborative Registry, a multi-center consortium of interdisciplinary BPD programs located in the United States and Sweden. In particular, this statistical analysis report approaches the problem based on not only baseline demographics and clinical diagnoses, but also detailed respiratory parameters at different postmenstrual ages (PMA). This enables prediction of need for earlier measures of intervention which posses significant clinical benefits to patients with the illness.

The data contains a set of demographic, diagnostic, and respiratory parameters of infants with sBPD admitted to collaborative NICUs and with known respiratory support parameters at 36 weeks and 44 weeks PMA. Detailed procedures and analysis are reported as the following sections.

We start by examine holistically missingness pattern in the data set and display variables that more than 20% of the case is missing. The reason for choosing 20% is not a strict threshold. We can see from the table 1 that all of the 44 week related measurements and 'any_surf' are those that selected. So we may considered drop those variables and only build the model based on 36 week measurements. Since for a variable having more than 20% of missingness, imputation methods may not generate stable and unbiased predictions to fill in the gap. Despite those variables, we do have other predictors that are having missingness and we considering using Multivariate Imputation by Chain Equation (MICE) to generate imputated data set for model training and testing.

Table 1: Variables with missing more than 20 Percent

	Missing_num	Missing_Pct
inspired_oxygen.44	451	45
p_delta.44	451	45
weight_today.44	449	45
peep_cm_h2o_modified.44	449	45
any_surf	433	43
ventilation_support_level_modified.44	427	43
med_ph.44	427	43

Exploratory Analysis

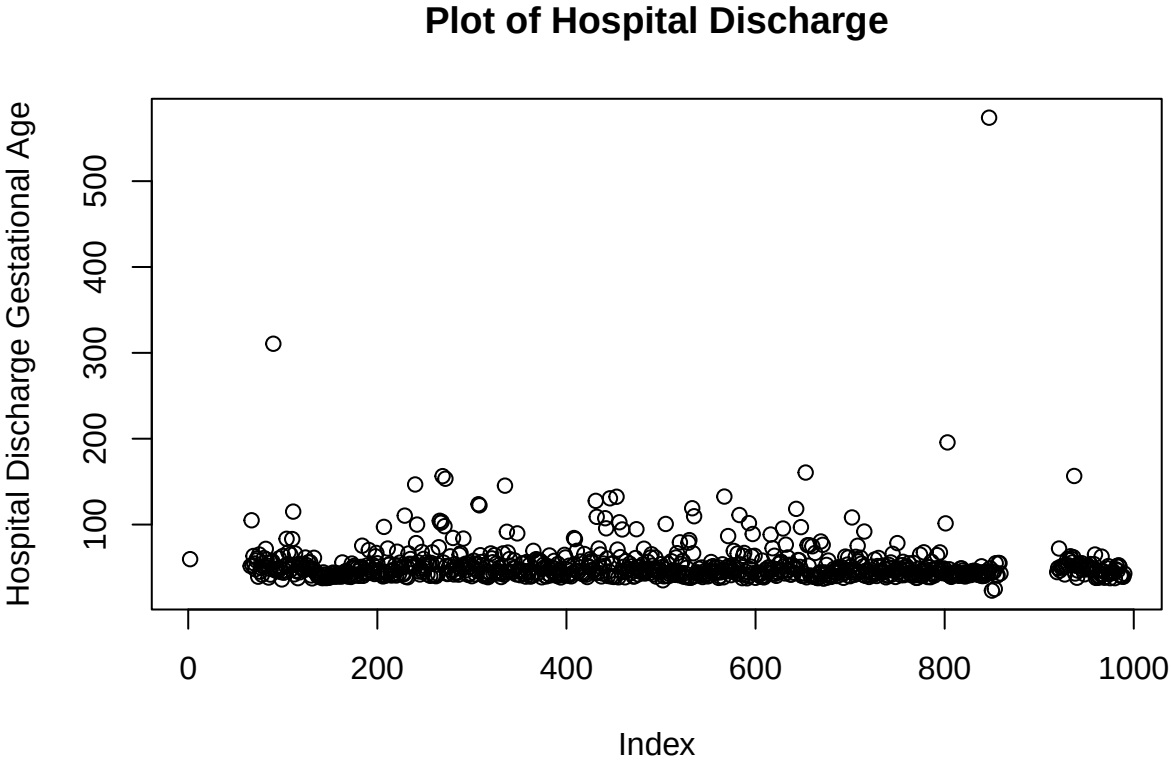
Then we conduct Exploratory Data Analysis to identify any irregular and meaningful patterns in the dataset. There is a observation that is repeated four times in the data (id = 2000824), and we should only kept one of its record. Then for the outcome, we are presented with two binary outcomes 'Death' and 'Trach', each

stands for death and tracheotomy placement. We decide to combine both into a composite outcome variable as 'res'. In the context, this is a binary variable meaning negative outcomes where 1 including dead or having tracheotomy placement, and 0 other wise. While combining, we discovered there are two observations whose 'Death' outcome is missing, case 879 and 191. And by examining their predictors, we decided to code 'Death' of 191 to be 'No', since it has record for a hospital discharge week. This may imply the patient not dead, and drop 879 since it does not have a valid hospital discharge week, we can not infer. Then we examining the 'center' variable. By looking at the distribution of center(its a multi-center study) from table 2, center 20 and 21 have very few cases, 4 and 1. We decided to drop those observations as their small sample will not provide valid and valuable predictions for incoming patients in those two centers if we are going to include center as one of the variables in the model.

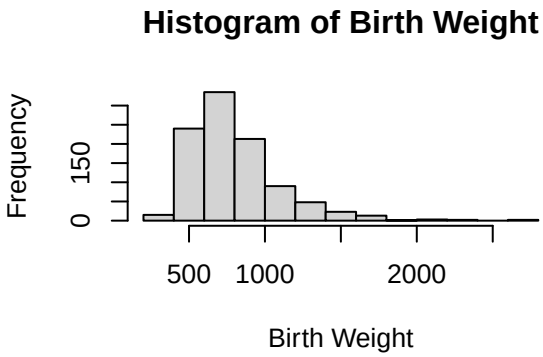
Table 2: Distribution of number of cases by Center

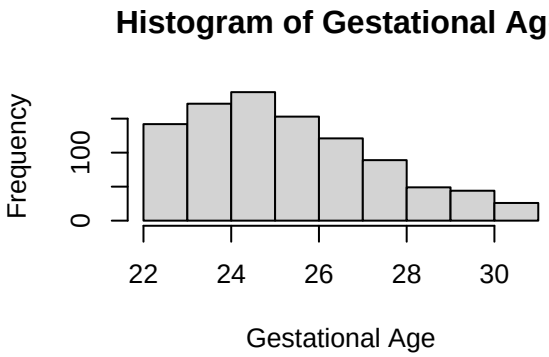
1	2	3	4	5	7	12	16	20	21
55	630	57	59	40	32	69	38	4	1

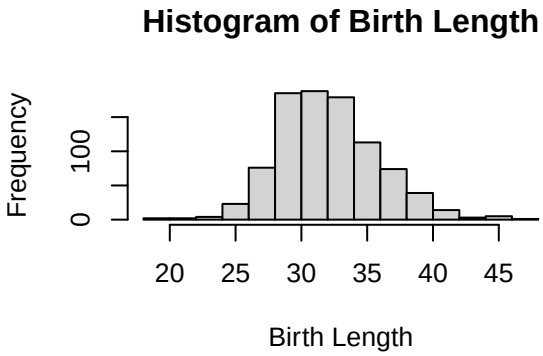
Next we start evalutaing outliers. For the importance of hospital discharge in the data, we plotted the variable and discovered there may be two outlier whose hospital dicharge weeks are bigger than 300. which is very far deviated from the most of the records. So we decided to drop these two cases since their presence may interfere our later model building process. In addition, we discover there are three patients who have a dischage week less then 36 recorded but also have multiple record for 36 week measurement in the dataset. We decided this may be error of recording in the data and removed those three observations.

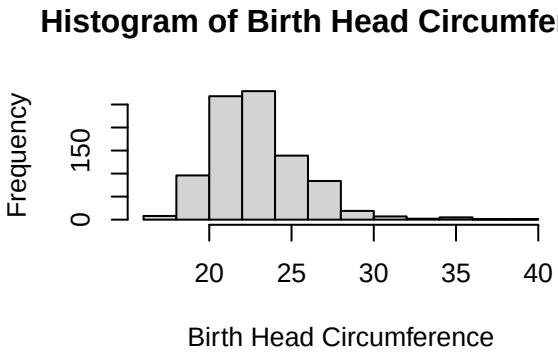


As we plotted the histogram of distribution for the four baseline numerical variables, we found out that birth weight and gestational age is a bit right-skewed. So we consider Logarithmic Transformation of this two variables when building the model









For multicollinearity, we calculated the VIF value of all the main effects from fitting a simple logistic regression. From the table 3 we discovered birth weight and ventilation support level at 36 weeks have high vif values meaning they are highly collinear to other predictors. So we may remove this two variables during the later stages.

Table 3: VIF Summary Table

Variable	VIF	Variable.cont	VIF.cont
center	2.619067	gender	1.157943
mat_race	1.757409	sga	2.855427
log(bw)	11.069557	any_surf	1.850707
log(ga)	5.960544	weight_today.36	3.006343
blength	5.071833	ventilation_support_level.36	26.961028
birth_hc	6.707449	inspired_oxygen.36	2.013262
del_method	1.417782	p_delta.36	6.390595
prenat_ster	1.000000	peep_cm_h2o_modified.36	6.116430
com_prenat_ster	1.179939	med_ph.36	1.329033
mat_chorio	1.314266	hosp_dc_ga	1.566693

Lastly, after finishing all the variable-specific checking, we are finalized with the dataset for model buidling. The following is a summary table of the remaining variables stratified by center. We want to see the significantly different variables among each centers and trying to identify potential interaction terms to add in the model building. By only displaying highly significantly different variables, we realize most of the numeric measurement variables are highly significant ones. This may be due to some systemic settings differing in each center, for example, different measuring equipment in terms of brand and versions. So we consider adding their interaction terms with the center while model building. So in conclusion, we decided the 'formula' of initial model to include all main effects except birth weight and ventilation support level, and interaction terms between center the remaining baseline measurement (ga,b_length,birth_hc).

Characteristic	Treatnment Center								p-value
	1, N = 55	**2**, N = 629	**3**, N = 54	**4**, N = 59	**5**, N = 40	**7**, N = 32	**12**, N = 68	**16**, N = 38	
mat_race									<0.001
0	20 / 38 (53%)	391 / 629 (62%)	30 / 49 (61%)	39 / 57 (68%)	12 / 40 (30%)	1 / 5 (20%)	16 / 68 (24%)	25 / 38 (66%)	
1	18 / 38 (47%)	191 / 629 (30%)	10 / 49 (20%)	16 / 57 (28%)	26 / 40 (65%)	3 / 5 (60%)	16 / 68 (24%)	5 / 38 (13%)	
2	0 / 38 (0%)	47 / 629 (7.5%)	9 / 49 (18%)	2 / 57 (3.5%)	2 / 40 (5.0%)	1 / 5 (20%)	36 / 68 (53%)	8 / 38 (21%)	
NA	17	0	5	2	0	27	0	0	
bw	685 (197)	833 (312)	774 (256)	830 (260)	605 (107)	725 (226)	780 (255)	889 (354)	<0.001
ga	25.65 (1.78)	25.87 (2.19)	25.74 (2.09)	25.69 (2.00)	24.08 (1.54)	25.09 (1.86)	26.07 (1.93)	26.29 (2.36)	<0.001
blength	30.8 (4.0)	32.8 (3.8)	32.4 (3.6)	33.2 (3.5)	29.5 (2.0)	32.1 (3.3)	32.4 (3.0)	33.7 (4.0)	<0.001
NA	8	24	0	1	1	6	36	0	
birth_hc	22.43 (2.07)	23.32 (2.73)	23.68 (3.21)	23.69 (2.89)	21.05 (1.51)	22.27 (2.22)	23.23 (2.79)	23.76 (2.92)	<0.001
NA	7	29	0	2	0	6	30	0	
com_prenat_ster	29 / 40 (73%)	414 / 524 (79%)	40 / 46 (87%)	24 / 42 (57%)	27 / 37 (73%)	15 / 23 (65%)	26 / 43 (60%)	26 / 33 (79%)	0.003
NA	15	105	8	17	3	9	25	5	
mat_chorio	14 / 29 (48%)	105 / 629 (17%)	4 / 32 (13%)	8 / 58 (14%)	14 / 39 (36%)	3 / 31 (9.7%)	5 / 68 (7.4%)	2 / 33 (6.1%)	<0.001
NA	26	0	22	1	1	1	0	5	
sga									0.008
Not SGA	33 / 54 (61%)	503 / 619 (81%)	40 / 51 (78%)	53 / 58 (91%)	32 / 40 (80%)	24 / 32 (75%)	51 / 68 (75%)	31 / 38 (82%)	
SGA	21 / 54 (39%)	116 / 619 (19%)	11 / 51 (22%)	5 / 58 (8.6%)	8 / 40 (20%)	8 / 32 (25%)	17 / 68 (25%)	7 / 38 (18%)	
NA	1	10	3	1	0	0	0	0	
any_surf									<0.001
Yes	31 / 55 (56%)	265 / 629 (42%)	51 / 54 (94%)	10 / 59 (17%)	33 / 40 (83%)	3 / 32 (9.4%)	47 / 68 (69%)	9 / 38 (24%)	
No	3 / 55 (5.5%)	70 / 629 (11%)	2 / 54 (3.7%)	5 / 59 (8.5%)	2 / 40 (5.0%)	3 / 32 (9.4%)	9 / 68 (13%)	7 / 38 (18%)	
Missing	21 / 55 (38%)	294 / 629 (47%)	1 / 54 (1.9%)	44 / 59 (75%)	5 / 40 (13%)	26 / 32 (81%)	12 / 68 (18%)	22 / 38 (58%)	
weight_today.36	2,073 (440)	2,137 (406)	2,131 (423)	2,127 (343)	1,922 (401)	2,169 (409)	2,044 (484)	2,220 (409)	0.011
NA	12	36	3	6	0	1	29	0	
ventilation_support_level.36									<0.001
0	7 / 55 (13%)	50 / 620 (8.1%)	5 / 53 (9.4%)	8 / 59 (14%)	0 / 40 (0%)	22 / 32 (69%)	1 / 49 (2.0%)	22 / 38 (58%)	
1	19 / 55 (35%)	425 / 620 (69%)	34 / 53 (64%)	34 / 59 (58%)	31 / 40 (78%)	8 / 32 (25%)	16 / 49 (33%)	14 / 38 (37%)	
2	29 / 55 (53%)	145 / 620 (23%)	14 / 53 (26%)	17 / 59 (29%)	9 / 40 (23%)	2 / 32 (6.3%)	32 / 49 (65%)	2 / 38 (5.3%)	
NA	0	9	1	0	0	0	19	0	
inspired_oxygen.36	0.63 (0.21)	0.32 (0.14)	0.31 (0.09)	0.40 (0.11)	0.36 (0.13)	0.36 (0.10)	0.40 (0.19)	0.35 (0.11)	<0.001
NA	14	36	2	3	0	1	29	0	
p_delta.36	7 (8)	5 (11)	7 (8)	5 (6)	4 (7)	0 (1)	9 (7)	1 (5)	<0.001
NA	17	38	5	15	13	1	32	0	
peep_cm_h2o_modified.36	7.4 (4.1)	6.5 (2.4)	7.7 (3.2)	5.6 (2.5)	8.8 (1.6)	1.7 (2.7)	6.6 (2.0)	3.3 (4.1)	<0.001
NA	18	40	9	6	0	1	34	0	
med_ph.36									<0.001
0	42 / 55 (76%)	596 / 620 (96%)	50 / 53 (94%)	48 / 59 (81%)	37 / 40 (93%)	30 / 32 (94%)	45 / 49 (92%)	34 / 38 (89%)	
1	13 / 55 (24%)	24 / 620 (3.9%)	3 / 53 (5.7%)	11 / 59 (19%)	3 / 40 (7.5%)	2 / 32 (6.3%)	4 / 49 (8.2%)	4 / 38 (11%)	
NA	0	9	1	0	0	0	19	0	
hosp_dc_ga	60 (NA)	53 (18)	47 (21)	NA (NA)	54 (18)	45 (7)	54 (14)	41 (3)	<0.001
NA	54	0	0	59	0	0	0	0	
res									<0.001
0	25 / 55 (45%)	546 / 629 (87%)	53 / 54 (98%)	47 / 58 (81%)	33 / 40 (83%)	31 / 32 (97%)	27 / 68 (40%)	37 / 38 (97%)	
1	30 / 55 (55%)	83 / 629 (13%)	1 / 54 (1.9%)	11 / 58 (19%)	7 / 40 (18%)	1 / 32 (3.1%)	41 / 68 (60%)	1 / 38 (2.6%)	
NA	0	0	0	1	0	0	0	0	

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

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

Variable Selection and Model Building

Before training model, we are considering performing variable selection so that we may come up with a sparse and concise model that are highly interpret able for early diagnostic prediction of the composite clinical outcome for the patients. The analysis aims using two methods and will validate the results of both methods in the model development stage. The two methods are Lasso and best subset, and both will incorporate cross validation for robustness and prevent over fitting. The reasons for choosing best subset rather than forward step wise regression are as follows: best subset ensures to find the best model by examining all possible combinations while step wise may not guarantee this by providing local optimal, and step wise may be subject to the ordering of predictors when dealing with many predictors. We have clearly more observations than the number of predictors so the over fitting problem of best subset may not occur. To overcome the computational burden, we found functions incorporate coordinate descent while searching (eg.l0learn) and also implement parallel computation both for lasso and best subset.

Before doing variable selection, we will split the data into train and test sets, and perform model processing on the train set. The preserved test sets will be saved and used for final validation after we acquire optimal combination of variables and models from both methods. With respect to the missing data, we will utilize technique of multiple imputation while doing variable selection. We preset for each imputation proportion of training and test set, and save each training and testing. The general scheme of variable selection is to perform cross validated variable selection methods on each of the imputated data set and combine those results into a final set as the variables that are selected.

For both methods, we are tuning different variables and utilize them differently for the purpose of variable selection. For lasso, since its penalized regularization will shrink certain coefficients of variables to be 0, we would see for each imputated data set, what are the variable coefficients its outputting with k-fold cross validation. We do not refit the lasso model after variable selection as it has been refitted in each cross validation.

On the other hand, best subset can be considered 'L0' penalty and we will also extract the coefficients generated by the minimal lambda value producing minimal cross validation errors. After averaging out the imputation results, we would then refit the variable selection on the full training set to obtain our final sets of coefficients for the best.subset model

Test train split

For Train and test data split, we incorporate the inbuild feature of MICE() function. It allows users to specificity the proportion of train-test split and will automatically do so for each of the imputation. We set the number of imputation to be five. So we will have 5 unique train and 5 unique test data sets. We will perform variable selection on each of the five train dataset and refit on the combined train dataset for final model coefficients. Lastly we shall evaluate our model on the combined test dataset, which the model building process has not seen.

Lasso

This is the result for lasso approach. In table 4, we can see the selected variables and their respective coefficients. The interpretation of the coefficients inline with the logistic regression. For a continuous predictor variable has a positive coefficient, it means that as the value of that predictor increases, the log-odds of the event happening (i.e., the probability of res outcome being 1, meaning bad outcome) also increases. And for categorical predictors variable has a positive coefficient, it means that being in the

particular level increases the log-odds of the event happening (i.e., the probability of res outcome being 1, meaning bad outcome) in comparison to reference level. And negative coefficients meaning the opposite. So for this lasso model, we can say that having Prenatal Corticosteroids and higher Fraction of Inspired Oxygen at 36 weeks are two example of positively associated predictors to the outcome, meaning patients with such traits are highly likely to develop bad outcomes (eg. Trachoestomy Placement or Death).

Table 4: Final Model for Lasso approach

Variable	Estimated Coefficients	Variable.cont	Estimated Coefficients.cont
(Intercept)	-5.04148	mat_chorioYes	0.22398
center 2	-1.73328	sgaSGA	0.17682
center 5	-0.33528	weight_today.36	-0.00070
center 7	-1.06956	inspired_oxygen.36	2.37830
center12	0.33180	p_delta.36	0.03956
center16	-1.89936	peep_cm_h2o_modified.36	0.00518
mat_race 1	0.00774	med_ph.36 1	0.11096
mat_race 2	0.02268	hosp_dc_ga	0.04644
blength	-0.00902	center 3:blength	-0.09676
birth_hc	0.08620	center 4:blength	-0.01230
prenat_sterYes	0.43470	center 5:birth_hc	-0.02420

Best Subset

This is the result for Best subset approach. In table 5, we can see variable selection results. And then we refit the set of coefficient to the whole train data, to obtain final coefficient estimates which is in table 6. The interpretation of the coefficients is very similar to the lasso interpretation, and inline with the logistic regression. For a continuous predictor variable has a positive coefficient, it means that as the value of that predictor increases, the log-odds of the event happening (i.e., the probability of res outcome being 1, meaning bad outcome) also increases. And for categorical predictors variable has a positive coefficient, it means that being in the particular level increases the log-odds of the event happening (i.e., the probability of res outcome being 1, meaning bad outcome) in comparison to reference level. And negative coefficients meaning the opposite. So for this bestsubset model, we identify two highly negatively associated predictors/levels: center 7 and 16. This means that the model predicts patients in center 7 and 16 are less likely to have the composite bad outcome.

Table 5: Variable Selection for Bestsubset approach

Variable	Estimated Coefficients
(Intercept)	-6.81998
center 4	-0.71750
center 5	-0.75164
center 7	-2.12438
center16	-5.23670
birth_hc	0.15822
prenat_sterYes	0.37920
weight_today.36	-0.00084
inspired_oxygen.36	2.16146
p_delta.36	0.03554
hosp_dc_ga	0.06180
center 5:log(ga)	-0.21170
center 2:blength	-0.07118
center 3:blength	-0.14884
center12:birth_hc	-0.01376

Table 6: Best Model for Bestsubset approach

Variable	Estimated Coefficients	Variable.cont	Estimated Coefficients.cont
(Intercept)	-4.5378472	ga	-0.0314878
center 2	-4.4048318	blength	-0.0672709
center 3	-7.5848738	center 2:ga	-0.0436005
center 4	5.3919719	center 3:ga	0.8375997
center 5	-2.8665928	center 4:ga	-0.1635432
center 7	-11.5397409	center 5:ga	0.0205511
center12	3.5240299	center 7:ga	0.0020910
center16	-8.9328084	center12:ga	-0.2114330
birth_hc	0.2132585	center 2:blength	0.0825248
prenat_sterYes	0.7042281	center 3:blength	-0.8047110
weight_today.36	-0.0012030	center 4:blength	-0.0756888
inspired_oxygen.36	2.5353185	center 5:blength	0.0136745
p_delta.36	0.0438233	center 7:blength	0.2621267
hosp_dc_ga	0.0646417	center12:blength	0.0499096

Model Evalutation

After acquiring both final models and their coefficients, we will evaluate the model on the test dataset. Since the models are logistic in nature, we propose the following model metrics as criteria for evaluation: 'AUC', 'Accuracy','Sensitivity','Specificity','Positive Predictive Value','Negative Predictive Value','F1', and also ROC curve for both models. Their meanings are as follows:

1. AUC (Area Under the ROC Curve):

- Meaning: AUC measures the model's ability to distinguish between the positive and negative classes across various probability thresholds. It represents the area under the Receiver Operating Characteristic (ROC) curve.

- Best Value: Higher values are better. A perfect classifier has an AUC of 1, while random guessing results in an AUC of 0.5.

2. Accuracy:

- Meaning: Accuracy is the proportion of correctly classified instances (both true positives and true negatives) out of the total.
- Best Value: Higher values are better. 100% accuracy means all predictions are correct.

3. Sensitivity (True Positive Rate):

- Meaning: Sensitivity measures the proportion of true positive predictions out of all actual positive instances. It indicates the model's ability to correctly identify positive cases.
- Best Value: Higher values are better, as you want to maximize the detection of positive cases. Sensitivity ranges from 0 to 1.

4. Specificity (True Negative Rate):

- Meaning: Specificity measures the proportion of true negative predictions out of all actual negative instances. It indicates the model's ability to correctly identify negative cases.
- Best Value: Higher values are better. Specificity ranges from 0 to 1.

5. Positive Predictive Value (Precision):

- Meaning: Precision is the proportion of true positive predictions out of all positive predictions made by the model. It measures the accuracy of positive predictions.
- est Value: Higher values are better. Precision ranges from 0 to 1.

6. Negative Predictive Value:

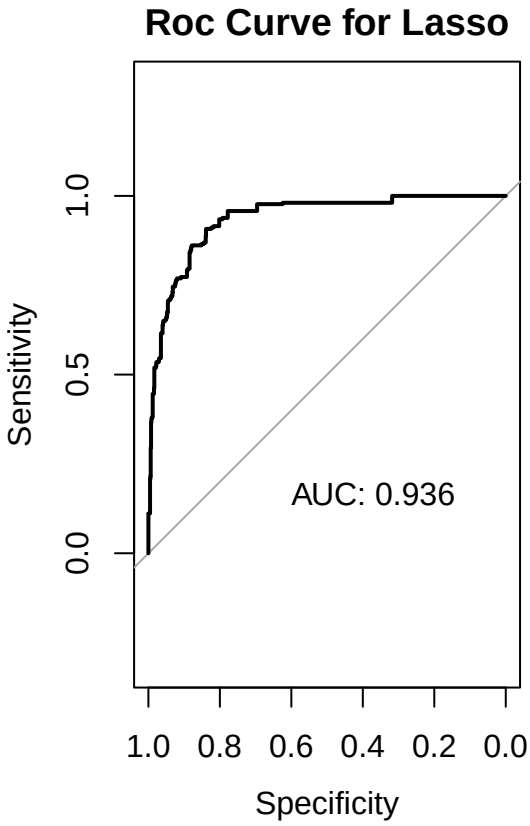
- Meaning: Negative Predictive Value is the proportion of true negative predictions out of all negative predictions made by the model.
- Best Value: Higher values are better.

7. F1-Score:

- Meaning: The F1-Score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is useful when there is an imbalance between the classes.
- Best Value: Higher values are better. The maximum F1-Score is 1, indicating perfect precision and recall.

In overall, the two models' performance are comparable across different metrics. They have very similar AUC and accuracy, and their ROC curve also look very align, indicating the relative robustness in predicting new cases. Best subset has higher Specificity and F-1 Score, meaning it is better at correctly identifying negative cases and reducing false alarms, which is especially important in scenarios where false positives are costly or undesirable. In the context of model, any false positive predictions may lead to unnecessary placement of tracheotomy. This can be devastating both biologically for the patient and economically for the family. And

may also lead to over-medication. So the final decision to choose the model between Lasso and Best subset can be very subjective and dependent on numerous factors, especially when their accuracy is close. It really boils down to the real-world scenario and application fields of such model, for example, ease of collection of data, quality of data and so on.



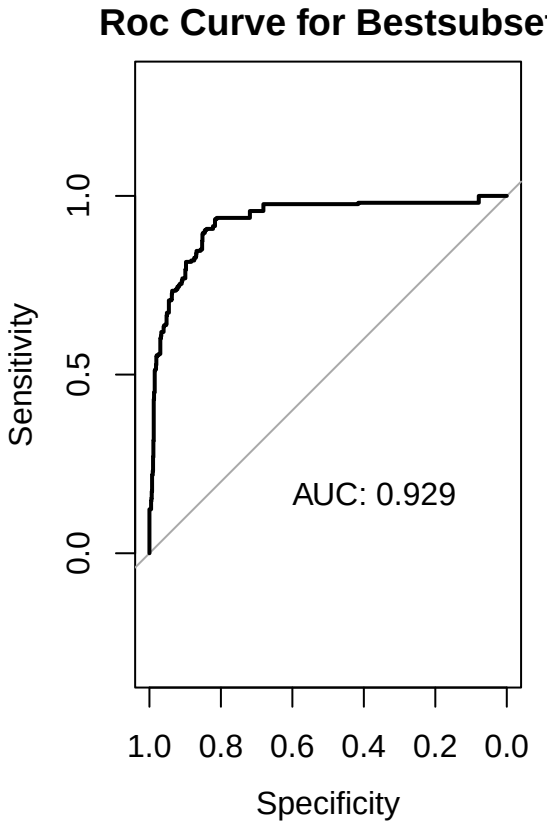


Table 7: Model Evaluation

	AUC	Accuracy	Sensitivity	Specificity	Positive Predictive Value	Negativc Predictive Value	F1
Lasso	0.9360227	0.8817844	0.4538462	0.9843318	0.8740741	0.8826446	0.5974684
Best Subset	0.9291634	0.8907063	0.5615385	0.9695853	0.8156425	0.9022298	0.6651481

Discussion and Limitation

This report outlines the step-by-step process of building regression models to predict a critical outcome: tracheostomy placement or patient mortality. The objective is to help determine need of placement for tracheostomy. We start by examining the data's characteristics and ensuring its quality by addressing missing information. We then explore the data, making transformations, removing unnecessary variables, and checking for unusual data points. Afterward, we select the most relevant variables, construct our models, and finally evaluate their performance.

In the end, we present two regression models—one created using the Lasso approach and another using the Best Subset approach. Both models perform well in terms of different performance matrics.

However, it's important to recognize that this study has limitations and unexplored aspects that could

further improve predictive accuracy and model applicability. While our focus has been on regression models, the problem we're addressing also has classification aspects. In future investigations, we could explore a broader range of machine learning methods, both supervised and unsupervised such as RandomForest, and we can apply ensemble learning to train the model. Within the realm of regression models, there's potential for applying multilevel mixed-effect models to account for the influence of different centers in our study. However, we didn't pursue this path due to unevenly distributed center data and limited sample size. Instead, we considered center as a categorical variable and its interactions with other predictors to capture the variability across various centers.

Moreover, our original goal was to predict outcomes using data from both 36 and 44 weeks. However, as we've shown, the absence of 44-week data made this infeasible. Future research could explore how our models perform with better data quality and advanced data imputation techniques, like Bayesian network learning.

Lastly, it's worth noting that both of our final models include a considerable number of predictors, which may not be necessary in every clinical prediction scenario, particularly in emergency care situations, as it will make interpretation harder. Achieving sparsity, or a simpler model with fewer predictors, is a potential goal. Techniques like integer risk models and categorizing variables could be explored to achieve this, although it may involve a trade-off between simplicity and predictive accuracy, and some valuable information could be lost.