Revised Project 2

Model to Predict The Need for Tracheostomy Placement in Neonates with Severe Bronchopulmonary dysplasia

ABSTRACT

Background: In recent past extended ventilation due to severe bronchopulmonary dysplasia (BPD) has risen as more infants survive severe BPD , making it the primary reason for tracheostomy in infants under the age of one year (Akangire & Manimtim, 2023). The need and the timing for tracheostomy placement has been an ongoing discussion with previous studies showing its benefits on growth (Akangire & Manimtim, 2023). The criteria indication in neonates with sBPD differs significantly among centers and their timings are currently unclear in pediatric care. The existing methods for predicting tracheostomy placement have shown an accurate prediction of the likelihood of tracheostomy placement based on baseline demographics and clinical diagnoses but lacks details on respiratory parameters and fail to provide predictions at various postmenstrual ages (PMA).This project aims to develop regression models that predicts the outcome of tracheostomy to inform the criteria indication for the need and optimal timing of tracheostomy placement among neonates.

Method: This project utilizes data collected from the BPD Collaborative Registry, where interdisciplinary BPD programs from the United States and Sweden collaborate to bridge knowledge gaps and improve care for children with severe BPD. Data were collected on respiratory parameters and demographic information at 36 and 44 weeks. The project utilizes a trained generalized linear mixed-effects regression approach for predicting the need for a tracheostomy placement. To enhance variable selection, Lasso and Ridge regression methods were compared. The variables selected by Lasso which included `prenat_steroids`, `ventilation_support_36wks and 44wks`,`inspired_oxy_36wks`, `peak_delta_44wks` were among the key respiratory variables that were used to develop the prediction model.The model was trained on 70% set and evaluated on 30% set of the same dataset. The model's performance were evaluated using metrics such as sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC).

Results: Data with 995 observations was used in developing the prediction model using data collected at two different time points (36weeks and 44weeks). On the test data, the models achieved an AUC of about 0.91 and 0.92 respectively demonstrating a sensitivity of 0.7456 and 0.7839 and specificity of 0.8678 and 0.8833 .The overall accuracy of the model on the test set was 0.91.

Conclusion: This prediction model is the first model for timing of tracheostomy placement in infants with sBDP that uses the respiratory parameters and considers the postmenstrual ages.The model based on 36 weeks postmenstrual age showed a better performance compared to the other model on 44 weeks. These findings shows the model's capacity to differentiate between positive and negative cases and shows its accuracy in identifying infants who truly need a tracheostomy and those whom immediate intervention might not be required. The model's overall accuracy further shows its reliability for informing clinical decision-making. Future work may involve evaluating the model using a different data source.

Introduction

Tracheostomy placement in infants with severe bronchopulmonary dysplasia (sBPD) is a critical intervention with great implications for both medical management and long term outcomes (Akangire & Manimtim, 2023). As advancements in neonatal care have led to increased survival rates among infants with severe

BPD, the need for prolonged ventilation has become a great indication for tracheostomy in this vulnerable population (Akangire & Manimtim, 2023). In pediatric healthcare, deciding when to do a tracheostomy in infants is crucial and the exact indication criteria and timing of tracheostomy placement in neonates with severe bronchopulmonary dysplasia (sBPD) remains a challenge. The decision making process surrounding the timing of tracheostomy placement remains complex, with ongoing debates regarding optimal criteria and potential benefits, particularly for infants with severe forms of BPD. Previous studies have shown the importance of early tracheostomy placement in influencing positive growth outcomes (Cheung & Napolitano, 2014). However, despite these considerations, precise predictive models to guide the optimal timing of tracheostomy in infants with severe BPD remain a critical area of research. Dr. Chris Schmid's project aims at solving the question surrounding when to perform tracheostomy in neonates with severe bronchopulmonary dysplasia (sBPD). Analyses done in the past have successful predicted the likelihood of a tracheostomy placement based on the baseline patient characteristics and clinical diagnoses but none have utilized the respiratory parameters which may provide accurate predictions for the need for tracheostomy during early postmenstrual ages (PMA) which could greatly help in advising on the timing of the procedure.

Methods

The data utilized in the analysis is sourced from a national database and comprises demographic details, diagnostic information, and respiratory parameters of infants diagnosed with sBPD. These infants were admitted to collaborative NICUs, and the dataset specifically includes data on their respiratory support status at 36 weeks and 44 weeks post menstrual age (PMA). The analysis used 995 records of data collected both at 36 and 44 weeks. Minimal data preprocessing was performed before the analysis. Duplicated records were removed, outlier points as observed in `hosp_dc_ga` were adjusted to a normal scale and variables were renamed to more meaningful naming. Multiple Imputation by Chained Equations (MICE) package in R was used to handle missing data. Binary variables were adjusted to the appropriate format to enhance consistency and interpretation of the data. Data for 36 weeks and 44 weeks were handled independently based on hospital gestational age variable.

Missing Data

Missingness were inspected for in each variable in the two datasets. There were notable differences in missingness between 36 and 44 weeks data as seen in Table 1 and Table 2

The observed missingness pattern at 36 weeks was likely MAR, indicating that missingness may have been influenced by other variables measured in the study. Contrary, the missingness at 44 weeks was likely missing at random (MAR). We also noted a pattern of variables that were missing together in some of the variables (i.e., `inspired_oxygen_44wks`, `peak_delta_44wks`, `weight_44`, `peep_cmH20_44`,`med_ph_44` and `vent_support_44`) which indicated that the absence of data in these variable was related to other measured covariates. `any_surf` had an abnormally different value for missingness at 44 weeks. We however noted that these variables were less correlated but there seem to be a fair correlation in the respiratory parameters as noted in the correlation Table 3

Based on Figure **??** for the distribution across centers, we consistently see lower percentages in Center 5 for maternal race, prenatal steroid use, and med_ph at 36 and 44 weeks. Conversely, Center 7 has lower gestational ages at birth and discharge, higher mortality rates, which indicates an increased need for tracheostomy. Additionally, Center 20 exhibits unique characteristics like lower birth weight. Noteworthy distributions involve birth weight, gestational age, ventilation support, inspired oxygen levels, and pH levels at different time points. These findings underlies the importance of considering center-specific factors in assessing the need for tracheostomy placement in neonates.

Table 1: Missing Data in the 36weeks Data

Variable	n	% Proportion
peak_delta_44wks	377	43.43
weight_44wks	376	43.32
inspired_oxygen_44wks	376	43.32
peep_cmH2o_mod_44wks	375	43.20
any_surf	363	41.82
vent_support_level_mod_44wks	357	41.13
med_ph_44wks	357	41.13
peak_delta_36wks	91	10.48
peep_cmH2o_mod_36wks	85	9.79
inspired_oxygen_36wks	72	8.29
weight_36wks	70	8.06
blength	68	7.83
birth_hc	66	7.60
comp_prenat_steroids	55	6.34
mat_race	34	3.92
mat_chorio	32	3.69
mat_ethn	31	3.57
prenat_steroids	30	3.46
ventilation_support_36wks	29	3.34
med_ph_36wks	29	3.34
sga	13	1.50
deliv_method	2	0.23
Death	1	0.12

Table 2: Missing Data in the 44weeks Data

Variable	n	% Proportion
any_surf	251	44.66
weight_44wks	90	16.01
inspired_oxygen_44wks	90	16.01
peak_delta_44wks	90	16.01
peep_cmH2o_mod_44wks	89	15.84
peak_delta_36wks	88	15.66
peep_cmH2o_mod_36wks	78	13.88
vent_support_level_mod_44wks	72	12.81
med_ph_44wks	72	12.81
inspired_oxygen_36wks	71	12.63
weight_36wks	69	12.28
blength	59	10.50
birth_hc	55	9.79
comp_prenat_steroids	43	7.65
ventilation_support_36wks	29	5.16
med_ph_36wks	29	5.16
prenat_steroids	24	4.27
mat_chorio	14	2.49
mat_race	13	2.31
mat_ethn	13	2.31
sga	8	1.42
deliv_method	2	0.36

Variable	N	Overall, N = 985	By Center										p-value
			1, N = 55	**2**, N = 629	**3**, N = 57	**4**, N = 60	**5**, N = 40	**7**, N = 32	**12**, N = 69	**16**, N = 38	**20**, N = 4	**21**, N = 1	
mat_race	933												
0		535 of 933 (57%)	20 of 38 (53%)	390 of 629 (62%)	31 of 52 (60%)	40 of 58 (69%)	12 of 40 (30%)	1 of 5 (20%)	16 of 69 (23%)	25 of 38 (66%)	0 of 3 (0%)	0 of 1 (0%)	
1		287 of 933 (31%)	18 of 38 (47%)	192 of 629 (31%)	10 of 52 (19%)	16 of 58 (28%)	26 of 40 (65%)	3 of 5 (60%)	16 of 69 (23%)	5 of 38 (13%)	1 of 3 (33%)	0 of 1 (0%)	
2		111 of 933 (12%)	0 of 38 (0%)	47 of 629 (7.5%)	11 of 52 (21%)	2 of 58 (3.4%)	2 of 40 (5.0%)	1 of 5 (20%)	37 of 69 (54%)	8 of 38 (21%)	2 of 3 (67%)	1 of 1 (100%)	
Unknown		52	17	0	5	2	0	27	0	0	1	0	
mat_ethn	932												
1		72 of 932 (7.7%)	4 of 34 (12%)	24 of 629 (3.8%)	14 of 55 (25%)	5 of 58 (8.6%)	8 of 40 (20%)	1 of 5 (20%)	8 of 69 (12%)	6 of 38 (16%)	1 of 3 (33%)	1 of 1 (100%)	
2		860 of 932 (92%)	30 of 34 (88%)	605 of 629 (96%)	41 of 55 (75%)	53 of 58 (91%)	32 of 40 (80%)	4 of 5 (80%)	61 of 69 (88%)	32 of 38 (84%)	2 of 3 (67%)	0 of 1 (0%)	
Unknown		53	21	0	2	2	0	27	0	0	1	0	
bw	985	807 (297)	685 (197)	832 (313)	765 (255)	833 (259)	605 (107)	725 (226)	781 (254)	889 (354)	1,089 (414)	590 (NA)	<0.001
ga	985	25.77 (2.14)	25.65 (1.78)	25.87 (2.19)	25.70 (2.08)	25.75 (2.03)	24.08 (1.54)	25.09 (1.86)	26.07 (1.91)	26.29 (2.36)	25.75 (2.50)	24.00 (NA)	<0.001
blength	908	32.5 (3.8)	30.8 (4.0)	32.7 (3.8)	32.2 (3.7)	33.2 (3.5)	29.5 (2.0)	32.1 (3.3)	32.4 (3.0)	33.7 (4.0)	32.5 (6.8)	29.0 (NA)	<0.001
Unknown		77	8	24	0	1	1	6	37	0	0	0	
birth_hc	910	23.19 (2.76)	22.43 (2.07)	23.31 (2.73)	23.46 (3.41)	23.73 (2.88)	21.05 (1.51)	22.27 (2.22)	23.23 (2.79)	23.76 (2.92)	24.01 (3.03)	21.00 (NA)	<0.001
Unknown		75	7	29	0	2	0	6	31	0	0	0	
deliv_method	983												
0		699 of 983 (71%)	40 of 55 (73%)	452 of 629 (72%)	40 of 57 (70%)	42 of 60 (70%)	26 of 40 (65%)	22 of 32 (69%)	49 of 67 (73%)	24 of 38 (63%)	3 of 4 (75%)	1 of 1 (100%)	
1		284 of 983 (29%)	15 of 55 (27%)	177 of 629 (28%)	17 of 57 (30%)	18 of 60 (30%)	14 of 40 (35%)	10 of 32 (31%)	18 of 67 (27%)	14 of 38 (37%)	1 of 4 (25%)	0 of 1 (0%)	
Unknown		2	0	0	0	0	0	0	2	0	0	0	
prenat_steroids	950												
0		126 of 950 (13%)	5 of 51 (9.8%)	85 of 628 (14%)	7 of 54 (13%)	12 of 59 (20%)	3 of 40 (7.5%)	4 of 30 (13%)	5 of 46 (11%)	4 of 37 (11%)	1 of 4 (25%)	0 of 1 (0%)	
1		824 of 950 (87%)	46 of 51 (90%)	543 of 628 (86%)	47 of 54 (87%)	47 of 59 (80%)	37 of 40 (92%)	26 of 30 (87%)	41 of 46 (89%)	33 of 37 (89%)	3 of 4 (75%)	1 of 1 (100%)	
Unknown		35	4	1	3	1	0	2	23	1	0	0	
comp_prenat_steroids	915												
0		311 of 915 (34%)	16 of 45 (36%)	195 of 609 (32%)	13 of 54 (24%)	30 of 55 (55%)	13 of 40 (32%)	12 of 27 (44%)	18 of 44 (41%)	11 of 37 (30%)	2 of 3 (67%)	1 of 1 (100%)	
1		604 of 915 (66%)	29 of 45 (64%)	414 of 609 (68%)	41 of 54 (76%)	25 of 55 (45%)	27 of 40 (68%)	15 of 27 (56%)	26 of 44 (59%)	26 of 37 (70%)	1 of 3 (33%)	0 of 1 (0%)	
Unknown		70	10	20	3	5	0	5	25	1	1	0	
mat_chorio	927	157 of 927 (17%)	14 of 29 (48%)	105 of 629 (17%)	4 of 34 (12%)	8 of 59 (14%)	14 of 39 (36%)	3 of 31 (9.7%)	6 of 69 (8.7%)	2 of 33 (6.1%)	1 of 3 (33%)	0 of 1 (0%)	
Unknown		58	26	0	23	1	1	1	0	5	1	0	
sex	985												
Ambiguous		3 of 985 (0.3%)	0 of 55 (0%)	2 of 629 (0.3%)	1 of 57 (1.8%)	0 of 60 (0%)	0 of 40 (0%)	0 of 32 (0%)	0 of 69 (0%)	0 of 38 (0%)	0 of 4 (0%)	0 of 1 (0%)	
Female		403 of 985 (41%)	21 of 55 (38%)	249 of 629 (40%)	22 of 57 (39%)	28 of 60 (47%)	17 of 40 (42%)	16 of 32 (50%)	28 of 69 (41%)	20 of 38 (53%)	1 of 4 (25%)	1 of 1 (100%)	
Male		579 of 985 (59%)	34 of 55 (62%)	378 of 629 (60%)	34 of 57 (60%)	32 of 60 (53%)	23 of 40 (57%)	16 of 32 (50%)	41 of 69 (59%)	18 of 38 (47%)	3 of 4 (75%)	0 of 1 (0%)	
sga	970												
0		772 of 970 (80%)	33 of 54 (61%)	502 of 619 (81%)	41 of 54 (76%)	54 of 59 (92%)	32 of 40 (80%)	24 of 32 (75%)	52 of 69 (75%)	31 of 38 (82%)	2 of 4 (50%)	1 of 1 (100%)	
1		198 of 970 (20%)	21 of 54 (39%)	117 of 619 (19%)	13 of 54 (24%)	5 of 59 (8.5%)	8 of 40 (20%)	8 of 32 (25%)	17 of 69 (25%)	7 of 38 (18%)	2 of 4 (50%)	0 of 1 (0%)	
Unknown		15	1	10	3	1	0	0	0	0	0	0	
any_surf	556												
0		101 of 556 (18%)	3 of 34 (8.8%)	70 of 334 (21%)	2 of 56 (3.6%)	5 of 16 (31%)	2 of 35 (5.7%)	3 of 6 (50%)	9 of 57 (16%)	7 of 16 (44%)	0 of 2 (0%)	0 of 0 (NA%)	
1		455 of 556 (82%)	31 of 34 (91%)	264 of 334 (79%)	54 of 56 (96%)	11 of 16 (69%)	33 of 35 (94%)	3 of 6 (50%)	48 of 57 (84%)	9 of 16 (56%)	2 of 2 (100%)	0 of 0 (NA%)	
Unknown		429	21	295	1	44	5	26	12	22	2	1	
weight_36wks	898	2,122 (413)	2,073 (440)	2,135 (408)	2,106 (425)	2,126 (339)	1,922 (401)	2,169 (409)	2,040 (479)	2,220 (409)	2,468 (173)	1,339 (NA)	0.003
Unknown		87	12	36	3	6	0	1	29	0	0	0	
ventilation_support_36wks	956												
0		116 of 956 (12%)	7 of 55 (13%)	50 of 620 (8.1%)	5 of 56 (8.9%)	8 of 60 (13%)	0 of 40 (0%)	22 of 32 (69%)	1 of 50 (2.0%)	22 of 38 (58%)	1 of 4 (25%)	0 of 1 (0%)	
1		585 of 956 (61%)	19 of 55 (35%)	424 of 620 (68%)	35 of 56 (62%)	34 of 60 (57%)	31 of 40 (78%)	8 of 32 (25%)	17 of 50 (34%)	14 of 38 (37%)	2 of 4 (50%)	1 of 1 (100%)	
2		255 of 956 (27%)	29 of 55 (53%)	146 of 620 (24%)	16 of 56 (29%)	18 of 60 (30%)	9 of 40 (22%)	2 of 32 (6.2%)	32 of 50 (64%)	2 of 38 (5.3%)	1 of 4 (25%)	0 of 1 (0%)	
Unknown		29	0	9	1	0	0	0	19	0	0	0	
inspired_oxygen_36wks	897	0.34 (0.15)	0.43 (0.21)	0.32 (0.14)	0.32 (0.10)	0.40 (0.12)	0.36 (0.13)	0.36 (0.10)	0.40 (0.19)	0.35 (0.11)	0.41 (0.28)	NA (NA)	<0.001
Unknown		88	14	36	2	3	0	1	29	0	2	1	
peak_delta_36wks	860	5 (10)	7 (8)	5 (11)	7 (8)	5 (6)	4 (7)	0 (1)	9 (7)	1 (5)	8 (9)	16 (NA)	<0.001
Unknown		125	17	39	7	15	13	1	32	0	1	0	
peep_cmH2o_mod_36wks	874	6.3 (2.9)	7.4 (4.4)	6.5 (2.4)	7.7 (3.1)	5.6 (2.5)	8.8 (1.6)	1.7 (2.7)	6.5 (2.0)	3.3 (4.1)	4.5 (3.1)	10.0 (NA)	<0.001
Unknown		111	18	41	11	6	0	1	34	0	0	0	
med_ph_36wks	956												
0		890 of 956 (93%)	42 of 55 (76%)	595 of 620 (96%)	53 of 56 (95%)	49 of 60 (82%)	37 of 40 (92%)	30 of 32 (94%)	46 of 50 (92%)	34 of 38 (89%)	4 of 4 (100%)	0 of 1 (0%)	
1		66 of 956 (6.9%)	13 of 55 (24%)	25 of 620 (4.0%)	3 of 56 (5.4%)	11 of 60 (18%)	3 of 40 (7.5%)	2 of 32 (6.2%)	4 of 50 (8.0%)	4 of 38 (11%)	0 of 4 (0%)	1 of 1 (100%)	
Unknown		29	0	9	1	0	0	0	19	0	0	0	
weight_44wks	541	3,645 (682)	3,612 (880)	3,704 (625)	3,644 (745)	NA (NA)	3,489 (622)	3,805 (733)	3,304 (768)	3,236 (987)	3,676 (662)	2,610 (NA)	0.013
Unknown		444	6	251	38	60	9	21	26	33	0	0	
vent_support_level_mod_44wks	562												
0		267 of 562 (48%)	9 of 51 (18%)	198 of 391 (51%)	12 of 20 (60%)	0 of 0 (NA%)	19 of 31 (61%)	10 of 12 (83%)	12 of 47 (26%)	5 of 5 (100%)	2 of 4 (50%)	0 of 1 (0%)	
1		142 of 562 (25%)	14 of 51 (27%)	97 of 391 (25%)	7 of 20 (35%)	0 of 0 (NA%)	9 of 31 (29%)	0 of 12 (0%)	13 of 47 (28%)	0 of 5 (0%)	1 of 4 (25%)	1 of 1 (100%)	
2		153 of 562 (27%)	28 of 51 (55%)	96 of 391 (25%)	1 of 20 (5.0%)	0 of 0 (NA%)	3 of 31 (9.7%)	2 of 12 (17%)	22 of 47 (47%)	0 of 5 (0%)	1 of 4 (25%)	0 of 1 (0%)	
Unknown		423	4	238	37	60	9	20	22	33	0	0	
inspired_oxygen_44wks	538	0.34 (0.15)	0.39 (0.20)	0.33 (0.13)	0.30 (0.11)	NA (NA)	0.30 (0.08)	0.37 (0.16)	0.41 (0.20)	0.27 (0.04)	0.35 (0.08)	0.21 (NA)	0.020
Unknown		447	9	252	38	60	9	21	25	33	0	0	
peak_delta_44wks	538	8 (14)	10 (11)	8 (16)	0 (2)	NA (NA)	1 (4)	0 (0)	8 (10)	0 (0)	8 (9)	0 (NA)	<0.001
Unknown		447	7	247	38	60	14	22	26	33	0	0	
peep_cmH2o_mod_44wks	540	4.2 (4.4)	8.5 (5.3)	3.8 (4.1)	2.9 (4.5)	NA (NA)	3.3 (4.3)	1.6 (3.8)	5.7 (4.2)	0.0 (0.0)	3.5 (4.0)	5.0 (NA)	<0.001
Unknown		445	7	249	39	60	9	20	28	33	0	0	
med_ph_44wks	562												
0		466 of 562 (83%)	25 of 51 (49%)	350 of 391 (90%)	19 of 20 (95%)	0 of 0 (NA%)	26 of 31 (84%)	8 of 12 (67%)	30 of 47 (64%)	4 of 5 (80%)	4 of 4 (100%)	0 of 1 (0%)	
1		96 of 562 (17%)	26 of 51 (51%)	41 of 391 (10%)	1 of 20 (5.0%)	0 of 0 (NA%)	5 of 31 (16%)	4 of 12 (33%)	17 of 47 (36%)	1 of 5 (20%)	0 of 4 (0%)	1 of 1 (100%)	
Unknown		423	4	238	37	60	9	20	22	33	0	0	
hosp_dc_ga	871	53 (27)	60 (NA)	53 (18)	55 (73)	NA (NA)	54 (18)	45 (7)	58 (34)	41 (3)	55 (13)	66 (NA)	<0.001
Unknown		114	54	0	0	60	0	0	0	0	0	0	
Tracheostomy	985												
0		843 of 985 (86%)	32 of 55 (58%)	565 of 629 (90%)	56 of 57 (98%)	49 of 60 (82%)	35 of 40 (88%)	31 of 32 (97%)	34 of 69 (49%)	37 of 38 (97%)	4 of 4 (100%)	0 of 1 (0%)	
1		142 of 985 (14%)	23 of 55 (42%)	64 of 629 (10%)	1 of 57 (1.8%)	11 of 60 (18%)	5 of 40 (12%)	1 of 32 (3.1%)	35 of 69 (51%)	1 of 38 (2.6%)	0 of 4 (0%)	1 of 1 (100%)	
Death	983												
0		929 of 983 (95%)	48 of 55 (87%)	599 of 628 (95%)	56 of 57 (98%)	58 of 59 (98%)	38 of 40 (95%)	32 of 32 (100%)	55 of 69 (80%)	38 of 38 (100%)	4 of 4 (100%)	1 of 1 (100%)	
1		54 of 983 (5.5%)	7 of 55 (13%)	29 of 628 (4.6%)	1 of 57 (1.8%)	1 of 59 (1.7%)	2 of 40 (5.0%)	0 of 32 (0%)	14 of 69 (20%)	0 of 38 (0%)	0 of 4 (0%)	0 of 1 (0%)	
Unknown		2	0	1	0	1	0	0	0	0	0	0	

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

[2] Kruskal-Wallis rank sum test

Table 3: Correlation Among Continous Variables with Missing Data

Variables 1	Variables 2	Correlation
weight_44wks	weight_36wks	0.7349777
birth_hc	blength	0.6938666
peak_delta_44wks	peak_delta_36wks	0.5939960
peak_delta_44wks	inspired_oxygen_36wks	0.5829899
peep_cmH2o_mod_44wks	peak_delta_44wks	0.5682687
peak_delta_36wks	inspired_oxygen_36wks	0.5528688
peak_delta_44wks	inspired_oxygen_44wks	0.4942900
inspired_oxygen_44wks	inspired_oxygen_36wks	0.4577585
peep_cmH2o_mod_44wks	inspired_oxygen_36wks	0.4457917
weight_36wks	blength	0.4427522
peep_cmH2o_mod_44wks	peak_delta_36wks	0.4414008
peep_cmH2o_mod_44wks	inspired_oxygen_44wks	0.4063490

Table 4: Summary Statistics for Continuous Variables

Variable	Mean	SD	Min	Max
bw	806.10	296.77	340.00	2725.0
ga	25.77	2.14	22.00	31.0
blength	32.49	3.82	18.00	48.0
birth_hc	23.19	2.76	13.50	38.3
weight_36wks	2120.90	413.58	710.00	3710.0
inspired_oxygen_36wks	0.34	0.15	0.21	1.0
peak_delta_36wks	5.27	9.74	0.00	46.0
peep_cmH2o_mod_36wks	6.33	2.91	0.00	18.0
weight_44wks	3646.12	682.09	3.00	5275.0
inspired_oxygen_44wks	0.34	0.15	0.21	1.0
peak_delta_44wks	7.62	14.19	0.00	52.0
peep_cmH2o_mod_44wks	4.30	4.46	0.00	20.0
hosp_dc_ga	52.78	26.53	23.10	573.9

Imputation Method For Missing Data

In addressing missing data, the Mice package was used for multiple imputations. Five complete datasets, each containing imputed values, were generated for variables relevant to both the 36 and 44 weeks datasets. Variables such as `mat_race`, `Death`, `record_id` and the `hosp-dc_ga` were intentionally excluded from the imputation process. The decision for the imputation was guided by the necessity to ensure a comprehensive analysis by addressing missing values in key variables. The variables that were included in the imputation were based on the initial exploratory analysis of missing data and the missingness patterns that were observed. This approach ensured completeness of the data for subsequent analyses.

Variable Selection

Variables used in the analysis were selected using the Lasso regression approaches to identify the most influential predictors in the prediction model. A 10-fold cross-validation procedure was employed for evaluating the performance of lasso regression models in each imputed dataset. Data were partitioned into ten folds, with nine folds being used for training and one for testing the model in each iteration over 10 iterations with each fold being used as the test set exactly once. Using the `cv.glmnet` function in R, lasso regressions models were applied to each imputed dataset, with lasso model being regularized using the L1 norm penalty and alpha set to 1, establishing a pure lasso model. Cross-validation served as a penalty to identify the optimal lambda value that minimized the cross-validation error, guiding the selection of coefficients to be used in

Table 5: Estimated Coefficients

	36wks Lasso	36wks Ridge	44wks Lasso	44wks Ridge
(Intercept)	-4.8861017	-5.0143979	-3.3177264	-4.3548789
mat_ethn2	-	0.0828487	-	-0.0400574
bw	-	0.0001963	8.09e-05	0.0002985
ga	-	0.0026431	-	0.0090615
blength	-	0.0185823	0.0077619	0.0180804
birth_hc	-	0.0144139	1.48e-05	0.0191505
deliv_method1	-	-0.1598759	-	-0.1391492
prenat_steroids1	0.4229124	0.5472254	0.3205985	0.4599446
comp_prenat_steroids1	0.3062693	0.3718882	0.2205585	0.3274740
mat_chorioYes	-	0.0023181	-0.0205712	-0.1138766
sexFemale	-	0.0458371	-	0.0110322
sexMale	-	-0.0354556	-	0.0032164
sga1	-	-0.1992338	-4.73e-05	-0.1569705
any_surf1	-0.0873298	-0.1565563	-0.2141751	-0.3084469
weight_36wks	-7.64e-05	-0.0003247	-8.57e-05	-0.0001154
ventilation_support_36wks1	-0.4112728	-0.5623810	-0.4235272	-0.4850421
ventilation_support_36wks2	0.4958344	0.5425511	0.1683507	0.4602093
inspired_oxygen_36wks	2.6486084	2.1886630	1.2560324	1.3146768
peak_delta_36wks	-	-0.0012517	-0.001096	-0.0042497
peep_cmH2o_mod_36wks	0.0012712	0.0277915	0.0101575	0.0260384
med_ph_36wks1	-0.030524	-0.1135016	-	-0.1803560
weight_44wks	-2.7e-05	-0.0000851	-1e-07	-0.0001247
vent_support_level_mod_44wks1	-	0.0861670	-	-0.0419239
vent_support_level_mod_44wks2	1.0004103	0.8940403	0.9708867	0.8844280
inspired_oxygen_44wks	-	0.0142531	0.1397648	0.6343505
peak_delta_44wks	0.0092389	0.0178546	0.0048664	0.0146608
peep_cmH2o_mod_44wks	0.1348571	0.1000388	0.0941361	0.0881279
med_ph_44wks1	0.0550687	0.2307804	0.0936857	0.3326471

the prediction model. An appropriate lambda in lasso shrinks variable coefficients to zero when there is no association with the outcome. The optimal lambda was determined for each imputed set, the models were refitted to each full imputed dataset, and the coefficients were extracted for subsequent prediction model. An aggregate for the coefficients from each imputed dataset were averaged to find the coefficients that were used in prediction modelling.

Lasso selects fewer predictors with non-zero coefficients from the the after 36 weeks data and more from weeks 44 data with minimal overlap. Both selects on `prenatal steroids`,`complete prenatal steroids` `maternal chorioamnionitis`, `ventilation support`, `peak delta`, `peep_cmH2o modification`, and `medication administration for pulmonary hypertension` predictors.These coefficients were included in developing the predictive models.

Model Derivation

Multilevel Logistic Regression Model

In the development of the prediction model for tracheostomy placement, the analysis centered on the imputed dataset obtained through the Multiple Imputation by Chained Equations (MICE) approach. Transforming the imputed dataset into a long format facilitated subsequent steps in model development. To evaluate the model's performance, we implemented a train-test split, allocating 70% of the data for training and the remaining 30% for testing. We employed a mixed effects logistic regression model utilizing the generalized logistic mixed-effects regression approach, with a random intercept for the center variable to account for the variability between centers. Variables drawn from Lasso regression models were included as fixed effects in both 36 and 44 weeks models. The 44weeks model included more predictors than the 36 weeks model (`birth_hc`, `any_surf`, `inspired_ocygen`). Significant interactions were also included in the model and adjusted for the variables that had significant interactions in the model such as `weight at 36weeks`. The model was trained on the training set and subsequent evaluation on the test set, offering predictions based on the optimal threshold of 0.2 and 0.3 for the probability of tracheostomy placement in 36 weeks and 44 weeks respectively.

The mixed effects model at 36 weeks is represented with the equation below. The model includes the fixed effects and random effects. The Fixed effects contains labeled, the values of β_i and u_j as the random intercept for the center variable to account for the variability between different centers.

The Model for the 36 weeks data is given by the equation below

$$\begin{aligned}\text{logit}(\Pr(\text{Trach} = 1)) =& \beta_0 + \beta_1 \times \text{weight\_36wks} \\ &+ \beta_2 \times \text{prenat\_steroids} + \beta_3 \times \text{comp\_prenat\_steroids} \\ &+ \beta_4 \times \text{ventilation\_support\_36wks} + \beta_5 \times \text{med\_ph\_44wks} \\ &+ \beta_6 \times \text{vent\_support\_level\_mod\_44wks} + \beta_7 \times \text{inspired\_oxygen\_36wks} \\ &+ \beta_8 \times \text{peep\_cmH2o\_mod\_44wks} \\ &+ \beta_9 \times \text{vent\_support\_level\_mod\_44wks} \times \text{inspired\_oxygen\_44wks} \\ &+ \beta_{10} \times \text{weight\_36wks} \times \text{vent\_support\_level\_36wks} \\ &+ \beta_{11} \times \text{weight\_36wks} \times \text{peep\_cmH20\_mod\_36wks} \\ &+ \beta_{12} \times \text{ventilation\_support\_36wks} \times \text{inspired\_oxygen\_36wks} \\ &+ u_{0\text{center}} + u_{1\text{center}}\end{aligned}$$

The Model for the 44 weeks data is given by the equation below

$$\begin{aligned}\text{logit}(\Pr(\text{Trach} = 1)) =& \beta_0 + \beta_1 \times \text{birth\_hc} \\ &+ \beta_2 \times \text{prenat\_steroids} + \beta_3 \times \text{ventilation\_support\_36wks} \\ &+ \beta_4 \times \text{vent\_support\_level\_mod\_44wks} + \beta_5 \times \text{inspired\_oxygen\_36wks} \\ &+ \beta_6 \times \text{peep\_cmH2o\_mod\_44wks} + \beta_7 \times \text{birth\_hc} \times \text{ventilation\_support\_36wks} \\ &+ \beta_8 \times \text{birth\_hc} \times \text{vent\_support\_level\_mod\_44wks} + \beta_9 \times \text{birth\_hc} \times \text{peep\_cmH2o\_mod\_44wks} \\ &+ \beta_{10} \times \text{vent\_support\_level\_mod\_44wks} \times \text{med\_ph\_44wks} + u_{0\text{center}} + u_{1\text{center}}\end{aligned}$$

Model Validation

Table 6: Measures of discrimination and calibration

Metric	Model for 44wks	Model for 36wks
Accuracy	0.8451537	0.8817204
AUC	0.8951035	0.9168237
Brier Score	0.1001370	0.0694576
F1 Score	0.6700252	0.6169154
Precision	0.5937500	0.5166667
Sensitivity	0.7687861	0.7654321
Specificity	0.8647845	0.8982456

On the test data, there were no major differences between the prediction model's for tracheostomy placement using 36 weeks and 44 data. The 36 weeks model achieved slightly higher AUC and accuracy compared to the 44 weeks model as indicated in Table 6. The prediction model using 36 weeks data is also seen to perform better than the 44 weeks data as seen in Figure 1 below.

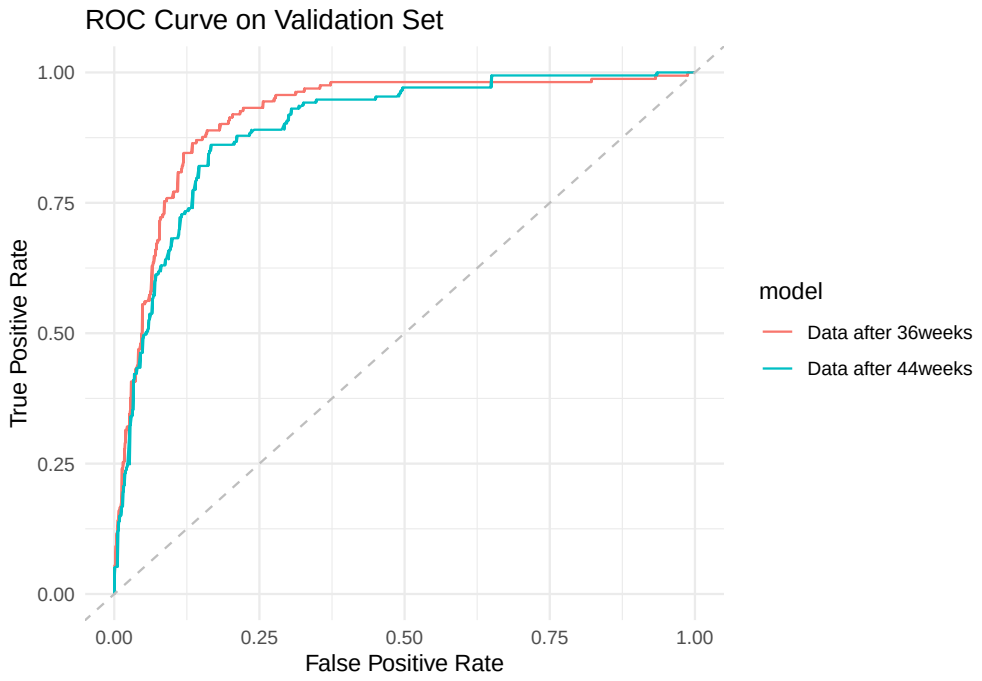

Figure 1: ROC Curve for Multilevel Logistics Regression Models

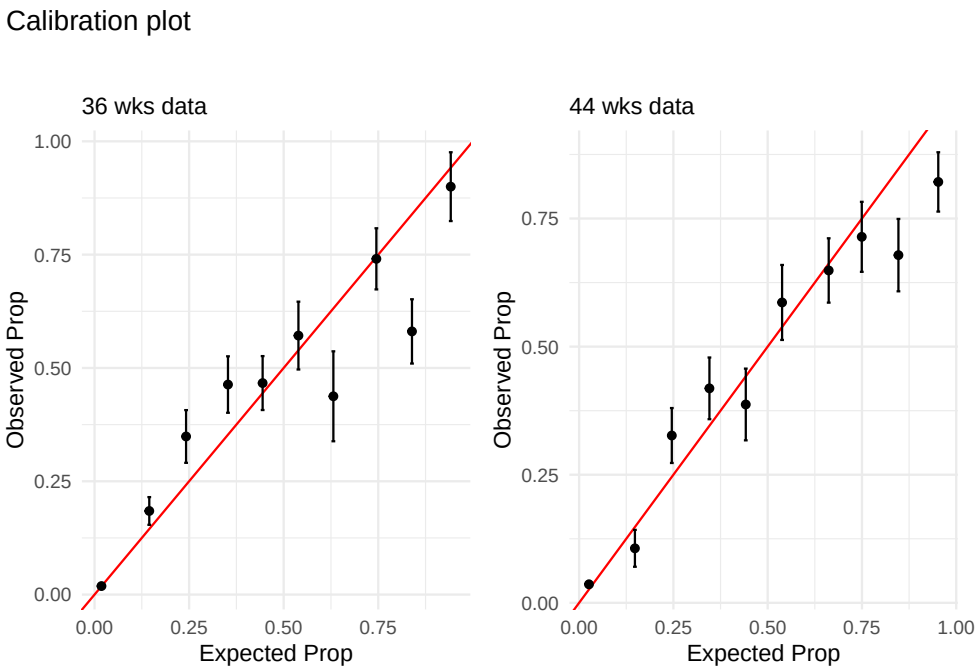

Figure 2: Calibration Plot for Multilevel Logistic Regression Model

Discussion

Interpretation of results

The prediction multilevel logistics model showed a balanced performance, with an AUC of 0.917 and 0.895, demonstrating a good discrimination ability. Sensitivity was measured at 0.7687 and 0.7469, suggesting that the models are effective in identifying infants who really needed tracheostomy placement with the model on 36weeks appearing to be more sensitive compared to the 44 weeks. The overall accuracy of the model on the test data was 0.85 again with the 36 weeks showing a better accuracy than the 44 weeks model, emphasizing the models reliability in making correct predictions. Also, the 36 weeks model had lower Brier score (0.06)compared to the 44 weeks model indicating the best calibration for the 36 weeks model. Based on the above, the 36 weeks multilevel logistics regression models show better performance than the 44 weeks model. Plots of the observed versus predicted values for each model showed that the line of best fit is quite accurate in predicting the timing and indication in the proportion of expected infants tracheostomy placement against the observed proportion of those requiring tracheostomy on average as seen in calibration Figure 2 with small standard errors. However there is still room for improvement as they do not perform optimally and maybe future work should consider looking at including more interaction terms . Overall, these findings showed the potential for adapting this predictive model to assist clinicians in making informed decisions about the timing of tracheostomy placement in infants, contributing to improved patient care and resource management.

From 44 weeks model, `prenatal steroids`,`ventilation support level at 44 weeks`, and `peep_cmH2o modification at 44 weeks` showed statistically significant estimates. Infants born to mothers who were on prenatal corticosteroids had 5.18 higher odds of requiring tracheostomy placement compared to thsoe whose mothers were not on any prenatal corticosteroids.

From the 36 weeks model `prenat_steroids`, `ventilation_support_36wks`, `inspired oxygen level at 36 weeks`, and `peep_cmH2o modification_36` showed statistically significant estimates. We also noted that, 1.53 higher odds of an infant requiring tracheostomy placement at 36 weeks for every unit increase in

`inspired_oxygen_36wks`. Contrary, each unit increase in `ventilation_support_36wks` reduced the odds of requiring traceostomy placement for an infant at 36 weeks by 0.90. These variables provides more emphasy on the predictors that may be considered in future works.

One strength in this project is the missingness was small for most of the variables more so the outcome variable which had only two observations missing.

Limitations

This project has quite a number of limitations. First the data was collected in different centers, with some of the centers having higher percentage of missing data. While there were no huge numbers of missingness, this limited our ability to develop predictive models for each center due to limited data. Also, there was limited observations for the two potential outcomes which may have limited our analysis. To help increase the sample size for the outcome, a composite outcome variable would have been constructed by combining `death` and `tracheostomy` but these outcomes seem to be very different and would limit the model's applicability/generalization. Additionally, missingness varied in the two time points when the data collected and this would have been as a result of the multicenter different data collection periods. Certain variables such as `any_surf` which may be strong predictors for the outcome had the largest percentage of missingness. There was an overlap in the predictors included in each model which made it alittle difficult to make the models distinctive in predicting the timing for tracheostomy placement among neonates with severe BDP besides the models seemed to be overfitting alittle and a more advanced method of variable selection would be recommended such as best subsets so as to get predictors that does not overfit the data and enhance better predictive models. Lastly, the model was built and trained on the same dataset and this may limit its generalizability to other populations. Additional future work will need to evaluate the model on an external dataset.

Conclusion

This project puts forth predictive models from infants with severe bronchopulmonary dysplasia from centers in the United States and Sweden. Based on lasso regression models, the measurements of several respiratory measures at two different time points allows for the prediction of the timing criteria for tracheostomy placements. These are among the first models that can help decide when a tracheostomy is required in infants with sBDP. The clinical application of these models is to aid in determining the optimal timing for tracheostomy placement in infants. We hope that the model's performance, in identifying infants who truly need a tracheostomy and those whom an immediate intervention may not be required, will be valuable tool for clinical decision making and management of infants which could potentially improve the care of infants needing tracheostomy placement and contribute to better healthcare outcomes and resource management.

References

1. Akangire G, Manimtim W. Tracheostomy in infants with severe bronchopulmonary dysplasia: A review. Front Pediatr. 2023 Jan 12;10:1066367. doi: 10.3389/fped.2022.1066367. PMID: 36714650; PMCID: PMC9878282

2. Cheung, N. H., & Napolitano, L. M. (2014). Respiratory Care. Respiratory Care, 59(6), 895-919. https://doi.org/10.4187/respcare.02971

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE,
                      message = FALSE,
                      warning = FALSE,
                      error = FALSE,
                      fig.align="center",
                      fig.pos = "H")


# load library
library(mice)
library(gtsummary)
library(tidyverse)
library(glmnet)
library(leaps)
library(kableExtra)
library(knitr)
library(pROC)
library(bestglm)
library(latex2exp)
library(ggplot2)
library(tableone)
library(Matrix)
library(MASS)
library(tidyr)
library(naniar)
library(dplyr)
library(lattice)
library(reshape2)
library(formatR)
library(lme4)
library(caret)
library(patchwork)
library(cowplot)

# load the data
trach_df = read.csv("project2.csv")

#check for duplicated IDs
dup_ids <- duplicated(trach_df$record_id) | duplicated(trach_df$record_id, fromLast = TRUE)

# rows with duplicated IDs
dup_entries <- trach_df[dup_ids, ] #789, 790, 791, 792

# blank values to NA
trach_df[trach_df == ""] <- NA

# keep only unique records
trach <- trach_df[!dup_ids, ] #995 30

# Select variables and rename
trach <- trach %>% #dplyr::select(-record_id, -center) %>%
```

```
  rename(prenat_steroids = prenat_ster,
         comp_prenat_steroids = com_prenat_ster,
         sex = gender, weight_36wks = weight_today.36,
         peep_cmH2o_mod_36wks= peep_cm_h2o_modified.36,
         weight_44wks = weight_today.44,med_PH_44wks = med_ph.44,
         peep_cmH2o_mod_36wks = peep_cm_h2o_modified.36,
         deliv_method = del_method,
         ventilation_support_36wks = ventilation_support_level.36,
         inspired_oxygen_36wks = inspired_oxygen.36,
         peak_delta_36wks = p_delta.36, med_ph_36wks = med_ph.36,
         vent_support_level_mod_44wks=ventilation_support_level_modified.44,
         peep_cmH2o_mod_44wks = peep_cm_h2o_modified.44,
         peak_delta_44wks = p_delta.44, med_ph_44wks = med_ph.44,
         Tracheostomy = Trach, inspired_oxygen_44wks = inspired_oxygen.44)

trach$deliv_method <- case_when(trach$deliv_method == 1 ~ 1,
                                trach$deliv_method == 2 ~ 0)

trach$prenat_steroids <- ifelse(trach$prenat_steroids == "No", 0,
                         ifelse(trach$prenat_steroids == "Yes", 1,
                                trach$prenat_steroids))

trach$comp_prenat_steroids <- ifelse(trach$comp_prenat_steroids == "No", 0,
                         ifelse(trach$comp_prenat_steroids == "Yes", 1,
                                trach$comp_prenat_steroids))

# make comp_prenat_steroids == No if prenat_steroids == No
trach$comp_prenat_steroids <- ifelse(trach$prenat_steroids == 0, 0,
                                     trach$comp_prenat_steroids)

trach$sex <- case_when(trach$sex == 1 ~ "1",
                       trach$sex == 2 ~ "0",
                       is.na(trach$sex) ~ "Ambiguous",
                       TRUE ~ as.character(trach$sex))

trach$sga <- ifelse(trach$sga == "Not SGA", 0,
                    ifelse(trach$sga == "SGA", 1, trach$sga))

trach$any_surf <- ifelse(trach$any_surf == "No", 0,
                           ifelse(trach$any_surf == "Yes", 1,
                                  trach$any_surf))

# create a composite outcome variable from Death and Tracheostomy
trach$Death <- ifelse(trach$Death == "No", 0,
                         ifelse(trach$Death == "Yes", 1,
                                trach$Death))

trach$Tracheostomy <- ifelse(trach$Tracheostomy == "No", 0,
                        ifelse(trach$Tracheostomy == "Yes", 1,
                                trach$Tracheostomy))

#trach$comp_outcome <- with(trach,
                    #   ifelse(Death == 1 | Tracheostomy ==
```

```
                        #            1,1, 0))

#trach$comp_outcome <- factor(trach$comp_outcome, levels = c("1", "0"))

#trach$comp_outcome <- as.integer(trach$comp_outcome)
trach$deliv_method <- as.integer(trach$deliv_method)

# All numeric
trach$bw <- as.numeric(trach$bw)
trach$blength <- as.numeric(trach$blength)
trach$weight_36wks <- as.numeric(trach$weight_36wks)
trach$weight_44wks <- as.numeric(trach$weight_44wks)
trach$peep_cmH2o_mod_36wks <- as.numeric(trach$peep_cmH2o_mod_36wks)
trach$peep_cmH2o_mod_44wks <- as.numeric(trach$peep_cmH2o_mod_44wks)
trach$ga <- as.numeric(trach$ga)


# All factors
fact <- function(data) {
    categorical_vars <- sapply(data, function(x) is.character(x) || is.integer(x))
    data[categorical_vars] <- lapply(data[categorical_vars], as.factor)

  return(data)
}
trach <- fact(trach)

# Filter data for after 36 weeks
trach_36wks <- trach %>%
  filter(hosp_dc_ga >= 36)

# Filter data to only after 44 weeks
trach_44wks <- trach %>%
  filter(hosp_dc_ga >= 44)


# Filter data for after 36 weeks
trach_36wks <- trach %>%
  filter(hosp_dc_ga >= 36)

# Distribution of Missing Data at 36 weeks
missing_df <- data.frame(
  Variable = names(trach_36wks),
  missing_count = sapply(trach_36wks, function(x) sum(is.na(x))))

missing_df$percent_missing <- round(missing_df$missing_count / nrow(trach_36wks) * 100, 2)

missing_df <- missing_df %>%
  arrange(desc(percent_missing))

# Remove row names
rownames(missing_df) <- NULL

# Select only those with missing records
```

```r
missing_df <- missing_df %>%
  filter(missing_count > 0)

missing_df$missing_count <- round(missing_df$missing_count, 2)
missing_df$percent_missing <- round(missing_df$percent_missing, 2)

  missing_df %>%
  mutate() %>%
  mutate_all(linebreak) %>%
  kbl(caption = "Missing Data in the 36weeks Data",
  col.names = linebreak(c("Variable", "n", "% Proportion")),
  booktabs = T, escape = T, align = "c") %>%
  kable_styling(latex_options = c('hold_position'),
                font_size = 8)

# Filter data to only after 44 weeks
trach_44wks <- trach %>%
  filter(hosp_dc_ga >= 44)

# Distribution of Missing Data at 44 weeks
missing_df2 <- data.frame(
  Variable = names(trach_44wks),
  missing_count = sapply(trach_44wks, function(x) sum(is.na(x))))

missing_df2$percent_missing <- round(missing_df2$missing_count / nrow(trach_44wks) * 100, 2)

missing_df2 <- missing_df2 %>%
  arrange(desc(percent_missing))

# Remove row names
rownames(missing_df2) <- NULL

# Select only those with missing records
missing_df2 <- missing_df2 %>%
  filter(missing_count > 0)

missing_df2$missing_count <- round(missing_df2$missing_count, 2)
missing_df2$percent_missing <- round(missing_df2$percent_missing, 2)

  missing_df2 %>%
  #  mutate(n = nrow(trach_44wks)) %>%
  mutate() %>%
  mutate_all(linebreak) %>%
  kbl(caption = "Missing Data in the 44weeks Data",
  col.names = linebreak(c("Variable", "n", "% Proportion")),
  booktabs = T, escape = T, align = "c") %>%
  kable_styling(latex_options = c('hold_position'),
                font_size = 9)

# Distribution of variables by center
trach %>%
  dplyr::select(-c(record_id)) %>%
  tbl_summary(by = center,
```

```
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} of {N} ({p}%)")) %>%
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 2)) %>%
  add_overall() %>%
  add_n() %>%
  modify_header(label ~ "**Variable**") %>%
  modify_spanning_header(c("stat_1", "stat_2") ~ "**By Center**") %>%
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE, escape = F) %>%
  kable_styling(latex_options = c("scale_down"),
                font_size = 8)
# Check to see how correlated the missing data
var_missn <- trach[, colSums(is.na(trach)) > 0]
num_data <- var_missn[sapply(var_missn, is.numeric)]
corr_mat <- cor(num_data)

# Compute correlation matrix
cor_matrix <- cor(num_data, use = "complete.obs")

# Melt the correlation matrix for ggplot
cor_df <- melt(cor_matrix)

# Corr Matrix for missing data
ggplot(data = cor_df, aes(x = Var1, y = Var2)) +
  geom_tile(aes(fill = value), color = "white") +
  geom_text(aes(label = round(value, 1)), size = 2) +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space =
                         "Lab", name="Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1,
                                   size = 10, hjust = 1),
        axis.text.y = element_text(size = 10)) +
  coord_fixed() +
  labs(x = NULL, y = NULL, title=
         "Correlation Matrix for Missing")
var_missn <- trach[, colSums(is.na(trach)) > 0]
num_data <- names(var_missn)[sapply(var_missn, is.numeric)]
corr_mat <- cor(trach[, num_data], use = "pairwise.complete.obs")

# Remove diagonal and redundant values
corr_mat[!lower.tri(corr_mat)] <- NA
cor_df <- data.frame(corr_mat) %>%
  rownames_to_column() %>%
  gather(key = "variable", value = "correlation", -rowname) %>%
  filter(abs(correlation) > 0.40)

# Correlation greater than 0.4
cor_df %>%
  rename('Variable1' = 'rowname', 'Variable2' = 'variable') %>%
  filter(!(str_detect(Variable1, "_diff") | str_detect(Variable2, "_diff"))) %>%
  arrange(desc(correlation)) %>%
```

```
  kbl(caption = "Correlation Among Continous Variables with Missing Data",
  col.names = linebreak(c("Variables 1", "Variables 2", "Correlation")),
  booktabs = T, escape = T, align = "c") %>%
  kable_styling(latex_options = c('hold_position'),
                font_size = 8)

# Calculate summary statistics for numeric columns
summary_table <- trach %>%
  dplyr::select(variable = colnames(.)[sapply(., is.numeric)]) %>%
  summarise(
    variable = colnames(trach)[sapply(trach, is.numeric)],
    Mean = round(colMeans(., na.rm = TRUE), 2),
    SD = round(apply(., 2, sd, na.rm = TRUE), 2),
    Min = round(apply(., 2, min, na.rm = TRUE), 2),
    Max = round(apply(., 2, max, na.rm = TRUE), 2))

summary_table %>%
  mutate() %>%
  mutate_all(linebreak) %>%
  kbl(caption = "Summary Statistics for Continuous Variables",
  col.names = linebreak(c("Variable", "Mean", "SD",
                          "Min","Max")),
  booktabs = T, escape = T, align = "c") %>%
  kable_styling(latex_options = c('hold_position'),
                font_size = 8)

# Remove variables that should not be imputed
trach_sub36 <- trach_36wks[, !colnames(trach_36wks) %in% c("mat_race", "Death", "record_id",  "hosp_dc_
trach_df_mice_out <- mice::mice(trach_sub36, 5, pri=F)

#  Remove variables that should not be imputed in 44wks data
trach_sub44 <- trach_44wks[, !colnames(trach_44wks) %in% c("mat_race", "Death", "record_id", "hosp_dc_g
trach_df_mice_out2 <- mice::mice(trach_sub44, 5, pri=F)

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

# Store each imputed data set
trach_df_imp2 <- vector("list",5)
for (i in 1:5){
   trach_df_imp2[[i]] <- mice::complete(trach_df_mice_out2,i)
}

#########################################################
#### Lasso for 36 weeks data ####
#########################################################

set.seed(1)
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(Tracheostomy ~., data = df)[,-1]
  y.ord <- df$Tracheostomy

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

  # Lasso model with cross-validation
  lasso_mod_cv <- cv.glmnet(x.ord, y.ord, nfolds = 10,
                            foldid = folds, alpha = 1,
                            family = "binomial")

  # Get coefficients for the minimum lambda
  coefi <- coef(lasso_mod_cv, s = "lambda.min")


  return(coefi)
}

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


##########################################################
#### Ridge for 36weeks data ####
##########################################################
set.seed(2)
ridge <- function(df) {
  #' Runs 10-fold CV for ridge and returns corresponding coefficients
  #' @param df, data set
  #' @return coef, coefficients for minimum cv error

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

  # Generate folds
  k <- 10
  set.seed(2) # consistent seeds between imputed data sets
  folds <- sample(1:k, nrow(df), replace=TRUE)
```

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

  # Get coefficients
  coef <- coef(ridge_mod)
  return(coef)
}
# Find average ridge coefficients over imputed datasets
ridge_coef1 <- ridge(trach_df_imp[[1]])
ridge_coef2 <- ridge(trach_df_imp[[2]])
ridge_coef3 <- ridge(trach_df_imp[[3]])
ridge_coef4 <- ridge(trach_df_imp[[4]])
ridge_coef5 <- ridge(trach_df_imp[[5]])
ridge_coef <- cbind(ridge_coef1, ridge_coef2, ridge_coef3, ridge_coef4, ridge_coef5)
avg_coefs_ridge <- apply(ridge_coef, 1, mean)

#########################################################
#### Lasso for 44 weeks data ####
#########################################################

set.seed(3)
lasso2 <- 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(Tracheostomy ~., data = df)[,-1]
  y.ord <- df$Tracheostomy

  # Generate folds
  k <- 10
  set.seed(3) # 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$lambda.min)
  # Get coefficients
  coef <- coef(lasso_mod) #, lambda = lasso_mod$lambda.min)
  return(coef)
}

# Find average lasso coefficients over imputed datasets
lasso_coef1.2 <- lasso2(trach_df_imp2[[1]])
lasso_coef2.2 <- lasso2(trach_df_imp2[[2]])
lasso_coef3.2 <- lasso2(trach_df_imp2[[3]])
lasso_coef4.2 <- lasso2(trach_df_imp2[[4]])
```

```r
lasso_coef5.2 <- lasso2(trach_df_imp2[[5]])
lasso_coef2 <- cbind(lasso_coef1.2, lasso_coef2.2, lasso_coef3.2,
                     lasso_coef4.2, lasso_coef5.2)
avg_coefs_lasso2 <- apply(lasso_coef2, 1, mean)

#########################################################
#### Ridge for 44 weeks data ####
#########################################################
set.seed(4)
ridge2 <- function(df) {
  #' Runs 10-fold CV for ridge and returns corresponding coefficients
  #' @param df, data set
  #' @return coef, coefficients for minimum cv error

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

  # Generate folds
  k <- 10
  set.seed(4) # consistent seeds between imputed data sets
  folds <- sample(1:k, nrow(df), replace=TRUE)

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

  # Get coefficients
  coef <- coef(ridge_mod)
  return(coef)
}
# Find average ridge coefficients over imputed datasets
ridge_coef1.2 <- ridge2(trach_df_imp2[[1]])
ridge_coef2.2 <- ridge2(trach_df_imp2[[2]])
ridge_coef3.2 <- ridge2(trach_df_imp2[[3]])
ridge_coef4.2 <- ridge2(trach_df_imp2[[4]])
ridge_coef5.2 <- ridge2(trach_df_imp2[[5]])
ridge_coef2 <- cbind(ridge_coef1.2, ridge_coef2.2, ridge_coef3.2, ridge_coef4.2, ridge_coef5.2)
avg_coefs_ridge2 <- apply(ridge_coef2, 1, mean)

# table to show the coefficients from two models at 36 weeks and 44 weeks
mytable <- as.data.frame(cbind(
  "36wks Lasso" = avg_coefs_lasso,
  "36wks Ridge" = avg_coefs_ridge,
  "44wks Lasso" = avg_coefs_lasso2,
  "44wks Ridge" = avg_coefs_ridge2
  )
)

# Round to 5dp
mytable <- round(mytable, 7)
```

```r
variables_to_exclude <- c("center2","center3","center4","center5","center7","center12","center16","cente

# Filter out the rows for the specified variables
mytable <- mytable[!rownames(mytable) %in% variables_to_exclude, ]

mytable[mytable == 0] <- "-"

mytable %>%
  kbl(caption = "Estimated Coefficients", booktabs = TRUE) %>%
  kable_styling(latex_options = c("scale_down"),
                font_size = 8) #%>%
  #column_spec(1, bold = TRUE)


#########################################################
#### Model evaluation for lasso Ridge at 36 weeks ####
#########################################################
# roc and auc
#roc_lasso <- roc(trach_df_long$Tracheostomy,predict_probs1)
#roc_ridge <- roc(trach_df_long$Tracheostomy,predict_probs9)

#roc_curve_lasso <- data.frame(FPR = 1-roc_lasso$specificities,
 #                             TPR = roc_lasso$sensitivities)
#roc_curve_ridge <- data.frame(FPR = 1-roc_ridge$specificities,
 #                             TPR = roc_ridge$sensitivities)

#roc_data <- rbind(roc_curve_lasso,roc_curve_ridge)
#roc_data$model <- c(rep("Lasso",nrow(roc_curve_lasso)),
  #                  rep("Ridge",nrow(roc_curve_ridge)))

# plot ROC curves for models that were used in variable selection
#p36 <- ggplot(roc_data, aes(x = FPR, y = TPR, color = model)) +
 # geom_line() +
  #geom_abline(intercept = 0, slope = 1, color = "grey", linetype = "dashed") +
  #labs(x = "False Positive Rate", y = "True Positive Rate",
   #    title = "36 weeks") +
  #theme(plot.title = element_text(hjust = 0.5),
   #      plot.subtitle = element_text(hjust = 0.5,size = 8)) +
  #theme_minimal()

#########################################################
#### Model evaluation for lasso Ridge at 44 weeks ####
#########################################################
# roc and auc
#roc_lasso2 <- roc(trach_df_long2$Tracheostomy,predict_probs1.2)
#roc_ridge2 <- roc(trach_df_long2$Tracheostomy,predict_probs9.2)

#roc_curve_lasso2 <- data.frame(FPR = 1-roc_lasso$specificities,
 #                             TPR = roc_lasso$sensitivities)
#roc_curve_ridge2 <- data.frame(FPR = 1-roc_ridge$specificities,
 #                             TPR = roc_ridge$sensitivities)

#roc_data2 <- rbind(roc_curve_lasso2,roc_curve_ridge2)
```

```r
#roc_data2$model <- c(rep("Lasso",nrow(roc_curve_lasso2)),
 #                    rep("Ridge",nrow(roc_curve_ridge2)))

# plot ROC curves for models that were used in variable selection
#p44 <- ggplot(roc_data2, aes(x = FPR, y = TPR, color = model)) +
 # geom_line() +
  #geom_abline(intercept = 0, slope = 1, color = "grey", linetype = "dashed") +
  #labs(x = "False Positive Rate", y = "True Positive Rate",
   #    title = "44 weeks") +
  #theme(plot.title = element_text(hjust = 0.5),
   #      plot.subtitle = element_text(hjust = 0.5,size = 8)) +
  #theme_minimal()

# Arrange plots side by side
#calib_plot <- p36 + p44 +
 # plot_layout(guides = "collect") +
  #plot_annotation(title = "ROC Curves for Regression in Coefficients Selection")
#calib_plot
#####################################################################
#### DEVELOPING A PREDICTIVE MODEL using 36 weeks data ####
#####################################################################
trach_df_long <- mice::complete(trach_df_mice_out,action="long")
y <- trach_df_long$Tracheostomy

# partion data into train-test split
set.seed(2550)
train_indices <- createDataPartition(y, p = 0.7, list = FALSE)
train_data <- trach_df_long[train_indices, ]
test_data <- trach_df_long[-train_indices, ]

y_test <- as.factor(test_data$Tracheostomy)

# Fit a logistic model with random intercept
model2 <- glmer(Tracheostomy ~ bw + blength + deliv_method + prenat_steroids + comp_prenat_steroids + s
                  ventilation_support_36wks + peep_cmH2o_mod_36wks+ peep_cmH2o_mod_44wks+ med_ph_44wks
                  weight_36wks*ventilation_support_36wks + inspired_oxygen_44wks+ weight_36wks*peep_cmH
                  ventilation_support_36wks*weight_36wks +  (1|center), data = train_data,
                family = binomial(link = "logit"))

# Predict on test set
preds <- predict(model2, newdata = test_data, type = "response")

#####################################################################
#### DEVELOPING A PREDICTIVE MODEL using data after 44 weeks ####
#####################################################################
trach_df_long2 <- mice::complete(trach_df_mice_out2,action="long")
yy <- trach_df_long2$Tracheostomy

# train and test split
set.seed(2550)
train_idx <- createDataPartition(y, p = 0.7, list = FALSE)
train_df <- trach_df_long2[train_idx , ]
test_df <- trach_df_long2[-train_idx , ]
```

```r
yy_test <- as.factor(test_df$Tracheostomy)

# Fit a mixed-effects logistic regression model
model2.2 <- glmer(Tracheostomy ~ birth_hc + blength + prenat_steroids + comp_prenat_steroids + any_surf

preds_44 <- predict(model2.2, newdata = test_df, type="response")

# determine the threshold to use in 36 weeks data
roc_mod <- roc(predictor=preds,
               response=as.factor(y_test),
               levels = c(0,1), direction = "<")
plot(roc_mod, print.auc=TRUE, print.thres = TRUE)


roc_vals <- coords(roc=roc_mod, x = "all")
#head(roc_vals)

roc_vals[roc_vals$sensitivity > 0.75, ] %>% tail(n=1)

# determine the threshold to use with the 44 weeks data
roc_mod <- roc(predictor=preds_44,
               response=as.factor(yy_test),
               levels = c(0,1), direction = "<")
plot(roc_mod, print.auc=TRUE, print.thres = TRUE)


roc_vals <- coords(roc=roc_mod, x = "all")
#head(roc_vals)

roc_vals[roc_vals$sensitivity > 0.75, ] %>% tail(n=1)
# model evaluation indicators using the combined data weeks
# AUC, sensitivity, specificity, accuracy, precision and Brier score values of the fitted model
evaluation <- function(pred,y_test,threshold=0.2){
  #' get AUC, sensitivity, specificity, accuracy, precision and Brier scorevalues of the fitted model
  #' @param pred, the prediction values
  #' @param y.test, the labels of test dataset
  #' @param threshold, a numeric number of threshold to classify the probability to classes
  #' @return AUC, sensitivity, specificity, accuracy, and precision values
  #'
   pred_numeric <- as.numeric(as.character(pred))
  y_test_numeric <- as.numeric(as.character(y_test))

  # Check if the outcome is binary
  if (length(unique(y_test_numeric)) > 2) {
    stop("Response variable must be binary for ROC analysis.")
  }

  # ROC curve
  roc_object <- roc(y_test_numeric, pred_numeric, levels = levels(y_test), direction = "<")

  # AUC
  auc <- roc_object$auc
```

```r
  df <- data.frame(pred = as.numeric(pred_numeric > threshold), label = as.numeric(y_test_numeric))

  TP <- dim(df[(df$pred==1&df$label==1),])[1]
  TN <- dim(df[(df$pred==0&df$label==0),])[1]
  FP <- dim(df[(df$pred==1&df$label==0),])[1]
  FN <- dim(df[(df$pred==0&df$label==1),])[1]

  Recall = TP / (TP + FN)
  Precision = TP / (TP + FP)
  Brier_score = mean((pred_numeric - y_test_numeric)^2)
  F1_score = 2 * (Precision * Recall) / (Precision + Recall)

  return(c(AUC = auc, sensitivity = Recall ,
           specificity = TN / (TN + FP),
            accuracy = (TP + TN) / (TP + TN + FP + FN),
           precision = Precision,
           "F1 score" = F1_score,
            "Brier Score" = Brier_score))

  }
evaluation_metrics <- evaluation(preds, y_test, threshold = 0.2)

# model evaluation indicators for 44weeks
# AUC, sensitivity, specificity, accuracy, precision and Brier score values of the fitted model
evaln <- function(pred,yy_test,threshold=0.3){
  #' get AUC, sensitivity, specificity, accuracy, precision and Brier scorevalues of the fitted model
  #' @param pred, the prediction values
  #' @param y.test, the labels of test dataset
  #' @param threshold, a numeric number of threshold to classify the probability to classes
  #' @return AUC, sensitivity, specificity, accuracy, and precision values
  #'
   pred_numeric_44 <- as.numeric(as.character(pred))
  yy_test_numeric <- as.numeric(as.character(yy_test))

  # Check if response is binary
  if (length(unique(yy_test_numeric)) > 2) {
    stop("Response variable must be binary for ROC analysis.")
  }

  # ROC curve
  roc_obj <- roc(yy_test_numeric, pred_numeric_44, levels = levels(yy_test), direction = "<")

  # AUC
  auc <- roc_obj$auc

  df <- data.frame(pred = as.numeric(pred_numeric_44 > threshold), label = as.numeric(yy_test_numeric))

  TP <- dim(df[(df$pred==1&df$label==1),])[1]
  TN <- dim(df[(df$pred==0&df$label==0),])[1]
  FP <- dim(df[(df$pred==1&df$label==0),])[1]
  FN <- dim(df[(df$pred==0&df$label==1),])[1]

  recall = TP / (TP + FN)
```

```
  precision = TP / (TP + FP)
  brier_score = mean((pred_numeric_44 - yy_test_numeric)^2)
  f1_score = 2 * (precision * recall) / (precision + recall)

  return(c(AUC = auc, sensitivity = recall ,
           specificity = TN / (TN + FP),
            accuracy = (TP + TN) / (TP + TN + FP + FN),
           precision = precision,
           "F1 score" = f1_score,
            "Brier Score" = brier_score))

  }
evaluation_metric <- evaln(preds_44, yy_test, threshold = 0.3)

# Evaluation Metrics at 36 and 44 weeks
evaluation_36 <- evaluation(preds, y_test, threshold = 0.2)
evaluation_44 <- evaln(preds_44, yy_test, threshold = 0.3)

evaln_df <- data.frame(
  Metric = c("AUC", "Sensitivity", "Specificity", "Accuracy", "Precision",
             "F1 Score", "Brier Score"),Value = evaluation_metric)

eval_df <- data.frame(
  Metric = c("AUC", "Sensitivity", "Specificity", "Accuracy", "Precision","F1 Score",
             "Brier Score"),Value = evaluation_metrics)

# Combine the two data frames
merged_df <- merge(evaln_df, eval_df, by = "Metric", suffixes = c("_44wks", "_36wks"))

colnames(merged_df) <- c("Metric", "Model for 44wks", "Model for 36wks")

   merged_df%>%
  kable(caption = "Measures of discrimination and calibration", booktabs = TRUE, align = "c") %>%
  kable_styling(latex_options = c('HOLD_position'))

############################################################
#### evaluation of the model on full data and 44weeks ####
############################################################
# roc and auc
preds <- as.numeric(preds)
preds_44 <- as.numeric(preds_44)

logistic_36wks <- roc(test_data$Tracheostomy,preds)
logistic_44wks <- roc(test_df$Tracheostomy, preds_44)

roc_curve_36 <- data.frame(FPR = 1-logistic_36wks$specificities,
                           TPR = logistic_36wks$sensitivities)
roc_curve_44 <- data.frame(FPR = 1-logistic_44wks$specificities,
                           TPR = logistic_44wks$sensitivities)

roc_dat <- rbind(roc_curve_36,roc_curve_44)
roc_dat$model <- c(rep("Data after 36weeks",nrow(roc_curve_36)),
                   rep("Data after 44weeks",nrow(roc_curve_44)))
```

```r
# Store AUC values
auc_36wks <- round(logistic_36wks$auc, 3)
auc_44wks <- round(logistic_44wks$auc, 3)

# plot ROC curves for models that were used in variable selection
ggplot(roc_dat, aes(x = FPR, y = TPR, color = model)) +
  geom_line() +
  geom_abline(intercept = 0, slope = 1, color = "grey", linetype = "dashed") +
  labs(x = "False Positive Rate", y = "True Positive Rate",
       title = "ROC Curve on Validation Set") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5,size = 8)) +
  theme_minimal()

# calibration plot for logistic regression at 36weeks
num_cuts <- 10
calib_dat <-  data.frame(probs = preds,
                         bin = cut(preds, breaks = num_cuts),
                         class=
                         as.numeric(test_data$Tracheostomy)-1)
calib_dat <- calib_dat %>%
             group_by(bin) %>%
             dplyr::summarize(observed = sum(class)/n(),
                              expected = sum(probs)/n(),
                        se = sqrt(observed * (1- observed)/n()))

plot_36 <- ggplot(calib_dat) +
  geom_abline(intercept = 0, slope = 1, color="red") +
  geom_errorbar(aes(x = expected, ymin = observed - 0.8*se,
                    ymax = observed + 0.8*se),
                position = position_identity(),
                colour="black", width = .01)+
  geom_point(aes(x = expected, y = observed)) +
  labs(x="Expected Prop", y="Observed Prop",title = '',
    subtitle = "36 wks data")+
  theme_minimal()

# calibration plot for logistic regression at 44 weeks
num_cuts <- 10
calib_df <-  data.frame(probs = preds_44,
                         bin = cut(preds_44, breaks = num_cuts),
                         class = as.numeric(test_df$Tracheostomy)-1)
calib_df <- calib_df %>%
             group_by(bin) %>%
             dplyr::summarize(observed = sum(class)/n(),
                       expected = sum(probs)/n(),
                       se = sqrt(observed * (1- observed)/n()))
plot_44 <-ggplot(calib_df) +
  geom_abline(intercept = 0, slope = 1, color="red") +
  geom_errorbar(aes(x = expected, ymin = observed - 0.8*se,
                    ymax = observed + 0.8*se),
                position = position_identity(),
                colour="black", width = .01)+
```

```
  geom_point(aes(x = expected, y = observed)) +
  labs(x ="Expected Prop", y="Observed Prop", title = '',
       subtitle = "44 wks data") +
  theme_minimal()

# Arrange plots side by side
calib_plots <- plot_36 + plot_44 +
  plot_layout(guides = "collect") +
  plot_annotation(title = "Calibration plot")
calib_plots

#Multilevel Logistic Regression Models
```

It often involves a trade-off between sensitivity and specificity. If you want to increase sensitivity, you may need to lower the threshold, and if you want to increase specificity, you may need to raise the threshold.