Tracheostomy/Death Prediction Model in Neonates with Severe Bronchopulmonary Dysplasia at 36-week and 44-week Gestational Age

Han Ji

2023-12-16

Abstract

Background: Bronchopulmonary dysplasia (BPD) is a chronic lung condition affecting 10,000 to 15,000 premature infants annually in the United States. Despite advances in neonatal care, the number of severe BPD cases remain steady, particularly among extremely low birth weight (ELBW) infants (Kalikkot Thekkeveedu et al. 2017).

Methods: This study addresses the critical need for effective prediction models for composite BPD outcomes - the potential need for tracheostomy and death as well as timing. Based on the records of infants with severe BPD across the US and Sweden from BPD Collaborative Registry (n = 985 and n = 615 at 36 weeks and 44 weeks, respectively), we developed three predictor models at each timing, lasso regression with and without center and multilevel lasso regression with center as a random effect. Multiple imputation was performed to address missing data.

Results: The fitted models are evaluated on the validation set split from the original data, and the multilevel lasso model has the best performance overall(AUC = 0.810 and 0.855). Considering measurements at 44-week improved the performance of all three models. Our findings reveal significant variability in outcomes across different centers, underscoring the importance of considering clinical setting heterogeneity and timing. Meanwhile, the estimates agree that ventilation support, inspired oxygen, prenatal corticosteriod, and hospital discharge gestational age are the most significant predictors, consistent with the existing literature. However, limitations such as potential bias due to variable missing proportions across centers and the challenge of predicting outcomes for centers not represented in the training data were acknowledged. Also, the resulting model need to adjust for calibration.

Conclusion: In conclusion, this study provides a prediction model for BPD outcomes with easily accessible clinical measurements, and it accommodates the complexity of clinical settings and measurements at different timing. These results potentially lead to individualized care for premature infants with severe BPD at early stages.

Introduction

Bronchopulmonary dysplasia (BPD) is a chronic lung condition primarily affecting premature infants, marked by an imbalance between lung injury and repair during lung development (Kalikkot Thekkeveedu et al. 2017). As the most common respiratory morbidity in preterm infants, BPD impacts nearly 10,000 − 15000 neonates annually in the United States. Despite significant advancements in managing extremely low birth weight (ELBW) infants, the incidence of BPD over the past two decades hasn't decreased. The pathogenesis of BPD is complex and varied by genetic and epigenetic factors for different individuals. Gestational age and birth weight are two of the strongest predictors: the risk of BPD is directly proportional to the degree of prematurity and low birth weight. For instance, infants born at 23 weeks have a significantly higher incidence and severity of BPD compared to those born at 28 weeks. Chorioamnionitis is also a frequent cause of preterm births, which likely results in serious complications in mother and babies but also BPD (Jain et al. 2022).

On the other hand, there are life-saving interventions including oxygen supplementation and mechanical ventilation, but they can potentially disrupt lung development and cause damage for neonates, and thus mechanical ventilation should be avoid when possible (Kalikkot Thekkeveedu et al. 2017). To prevent this damage but also help infants acquire necessary oxygen, tracheostomy can be planted in patients with severe BPD when discharged from hospital (Paul 2023). Although tracheostomy is still hooked with a ventilator, it has benefits in infant lung and neural development including improving ventilator synchrony, weaning sedation requirements, etc. However, tracheostomy is associated with an increased risk of death, meaning whether using tracheostomy should be carefully considered.

Unal et al. conducted case studies with 9 neonates admitted to neonatal intensive care units (NICU), and they showed tracheotomy, if carried out correctly, makes infant nursing easier with low complication (Unal et al. 2015). Adaikalam et al. have developed a tracheostomy prediction model for neonatal BPD via lung and airway MRI, while MRI might not be available for neonates at many hospitals. Murthy et al. developed a prediction model to estimate the risk of death or tracheostomy in infants with severe BOD (Murthy et al. 2014). However this study didn't different centers nor measurements at different time points. At present, no prediction models for tracheostomy based on easily accessible clinical diagnosis have considered the homogeneity of centers and the measurements over time. This study aims to develop a tracheostomy/death prediction model in neonatal BPD with easily accessible measurements, and it also address timing consideration of whether the infant needs more care.

Methods

Study setting and population

Study participants were drawn from the BPD Collaborative Registry, a multi-center consortium of interdisciplinary BPD programs located in the United States and Sweden formed to address gaps in evidence and promote research to enhance the care of children with severe forms of BPD. The registry includes infants whose gestational age is less than 32 weeks and who have severe bronchopulmonary dysplasia (sBPD) (defined by 2001 NHLBI criteria; specifically, FiO2 ≥ 0.3 or positive pressure ventilation (invasive or non-invasive) at 36-weeks PMA). In the registry, standard demographic and clinical data are collected at four time points: birth, 36 weeks PMA, 44 weeks PMA and discharge (Paul 2023).

Between January 1 and July 19, 2021, a total of 999 records of neonates with BPD were in the registry. At the time of analysis, 10 BPD Collaborative centers had contributed data meeting study inclusion criteria. Of these, 3 were excluded due to duplicate records, 5 were excluded because the their center size (n = 4; n = 1) were too small in this study. The hospital discharge gestational age for these two centers were not lower compared to other centers overall, meaning the neonates in these centers didn't seem to be transferred to other centers. We removed these records because it may cause overfitting and problems in train-test split. Although these observations could be used in models that don't consider center, the resulting prediction results and accuracy could correspond to different sub-populations.

In addition, 2 records were excluded due to missing death indicators, and 4 were excluded due to outliers in gestational age at discharge. 985 observations were left for the following analysis. To consider timing of tracheostomy, we considered two different scenarios: predicting the status of infants given the measurements at 36 weeks (n = 985); predicting the status given the measurements at both 36 and 44 weeks. For the first scenario, we used all records without looking at the 44 weeks measurements. For the second scenario, we removed the observations discharged before 44 weeks. For the records with missing discharged weeks, we kept the ones with no more than 2 missing measurements at 44 weeks (n = 615).

Maternal race was excluded because its values in the data didn't match the codebook. The remaining variables have demographic information, supporting factors like steroids, measurements of infants at birth and at 36 weeks, tracheostomy usage, and death. We combined tracheostomy and death into a composite outcome - whether infants had tracheostomy or died - because the number of cases for each is too low. This means that our model aim to predict the BPD status of the infants about whether they would be likely to need tracheostomy at discharge or have risk to die. However, this model won't provided a relative risk between tracheostomy and death.

Model derivation

To perform variable selection and utilize a mixed effects model for different centers, Lasso regression models was chosen as a public available R package `glmmLasso` can construct mixed effects modes with L1 regularization (Groll 2023). There were no existing packages for mixed effects models for other variable selection methods like best subset, and thus we didn't consider it. For each timing scenario, two separate models were developed with center as a factor and without center. This modeling part was performed using `glmnet` package in R (Friedman et al. 2010). A multilevel lasso regression model was developed by treating the data clustered by center, and a random effect was fit for center while fixed effects for the rest. This modeling part was performed using `glmmLasso` package in R. One-hot encoding was applied to the train data, while center indicator was excluded from normalization for the multilevel lasso.

Imputation and cross-validation

Multiple imputation was applied following the guidance by He (He 2010). Specifically, we replaced the missing values for surfactant usage indicator because its missing proportion is too high. Also, we used the original outcome variables tracheostomy and death in imputation formula to retain the association between predictors and the outcome rather than the composite outcome. The data set was split into train set (n = 660) and validation set (n = 325) by 2:1 ratio sampled by center and the composite outcome. Multiple imputation with 5 imputed sets was done first in the train set, and the same imputation model was fit on the test set separately. This prevents the imputed values in the train set borrow information from test set and vice versa. For 36-week scenario, the variables related to 44-week measurements were excluded in model fitting. For 44-week scenario, another set of train-test split indices for the 44-week observations was created to obtain train and test set (n = 403 and n = 212 respectively) from the complete full set. All imputation was done by `mice` package in R (van Buuren and Groothuis-Oudshoorn 2011).

5-fold cross-validation was performed to estimate the optimal hyperparameter λ for all three models. The fold index was generated by center and remained the same. The cross-validation for two lasso regression models was done by function `cv.glmnet`. For the multilevel lasso regression, a grid search method was applied: for each validation set, AUC was calculated for 20 λ ranging evenly from 0 to the max possible value where all coefficients shrink to 0. The optimal λ was obtained by the highest average AUC. We then repeat this grid search within a smaller range bound the neighboring values of the optimal λ.

Model performance

Models were examined on the validation set based on their sensitivity, specificity, AUC, and F-score. The optimal threshold for prediction was determined by Youden Index, where the maximum of sensitivity + specificity was achieved (Ruopp et al. 2008). The final validation set was composed of 5 imputed sets together, which is equivalent to evaluating the models on each imputed set and taking the average.

Results

Study population characteristics

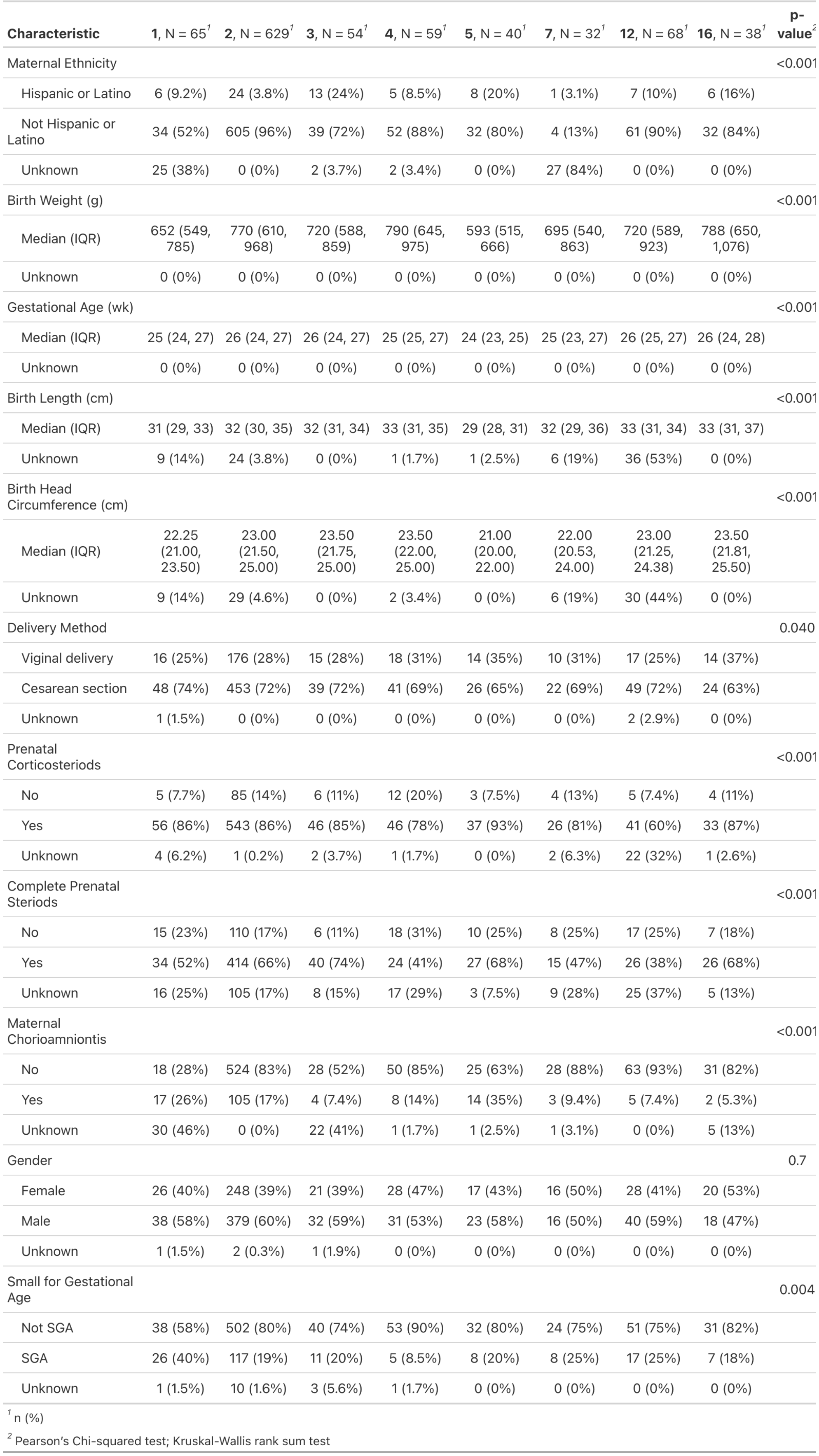
Figure 1-3 summarize the above characteristics stratified by center, and centers differ from each other significantly regarding the baseline and the outcome. Center 2 has the highest number of infants admitted (n = 630), while center like 16 only has 38. Infants in center 1 have a low median gestational age, short birth length, and birth weight, which is associated with the high proportion in composite outcome; however, infants in center 5 have worse baseline characteristics but has a much lower proportion in composite outcome. There is no obvious trend agreed across centers regarding the association between characteristics and the outcome.

Figure 4 shows the positive correlation between birth weight and gestational age, and it seems to have a regression line can determine whether the infant is small for gestational age (SGA) or not. This could imply multi-collinearity in the data since many clinical decisions and measurements are associated. In addition, most infants with composite outcome are not SGA, meaning they should have a better baseline health level but the results are counterintuitive.

Model Performance

The model coefficients are obtained by averaging the coefficients from each imputed sets for 36 and 44 weeks (Table 1, 2). For consistency, we calculate the intercept for each center by adding the intercept and the individual center intercept. We see the overall trend among coefficients is consistent across three models as the sign and the relative magnitude are comparable. No variables are selected out by multilevel lasso regression, but they are close to 0 if these variables are selected out by the lasso regression with or without center. We see the predictors prenatal corticosteriods and other 36-week measurements tend to have high positive estimates across three models. No predictors have high negative estimates except center intercepts.

The most significant difference among these three models comes from the center intercepts due to the model derivation. The lasso regression without using center as a predictor has the same intercept. The one uses center and the multilevel lasso model have varying intercepts, but we see the difference tends to be larger in the multilevel

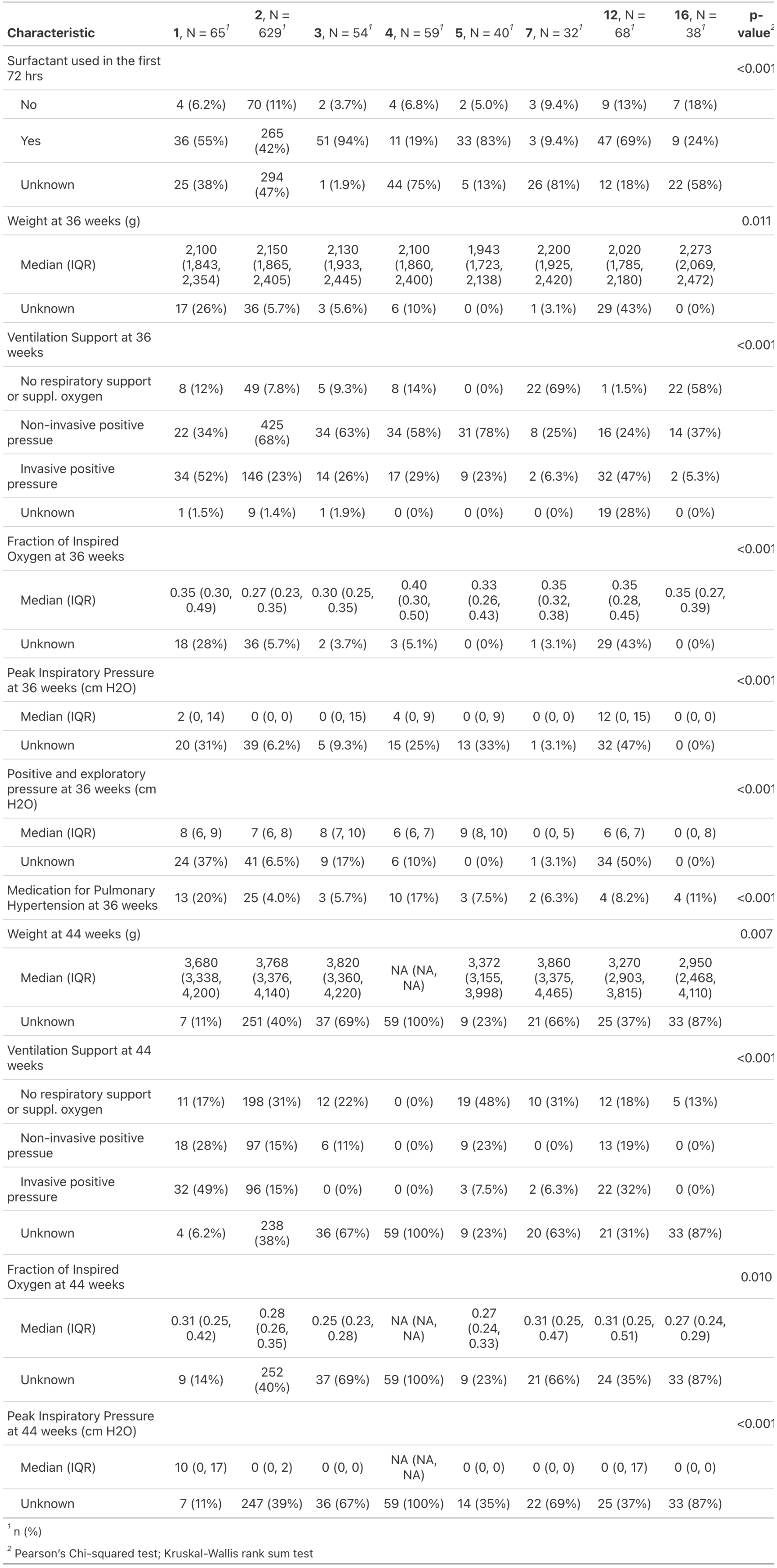


Characteristic	**1**, N = 65[1]	**2**, N = 629[1]	**3**, N = 54[1]	**4**, N = 59[1]	**5**, N = 40[1]	**7**, N = 32[1]	**12**, N = 68[1]	**16**, N = 38[1]	**p-value**[2]
Maternal Ethnicity									<0.001
Hispanic or Latino	6 (9.2%)	24 (3.8%)	13 (24%)	5 (8.5%)	8 (20%)	1 (3.1%)	7 (10%)	6 (16%)	
Not Hispanic or Latino	34 (52%)	605 (96%)	39 (72%)	52 (88%)	32 (80%)	4 (13%)	61 (90%)	32 (84%)	
Unknown	25 (38%)	0 (0%)	2 (3.7%)	2 (3.4%)	0 (0%)	27 (84%)	0 (0%)	0 (0%)	
Birth Weight (g)									<0.001
Median (IQR)	652 (549, 785)	770 (610, 968)	720 (588, 859)	790 (645, 975)	593 (515, 666)	695 (540, 863)	720 (589, 923)	788 (650, 1,076)	
Unknown	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Gestational Age (wk)									<0.001
Median (IQR)	25 (24, 27)	26 (24, 27)	26 (24, 27)	25 (25, 27)	24 (23, 25)	25 (23, 27)	26 (25, 27)	26 (24, 28)	
Unknown	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Birth Length (cm)									<0.001
Median (IQR)	31 (29, 33)	32 (30, 35)	32 (31, 34)	33 (31, 35)	29 (28, 31)	32 (29, 36)	33 (31, 34)	33 (31, 37)	
Unknown	9 (14%)	24 (3.8%)	0 (0%)	1 (1.7%)	1 (2.5%)	6 (19%)	36 (53%)	0 (0%)	
Birth Head Circumference (cm)									<0.001
Median (IQR)	22.25 (21.00, 23.50)	23.00 (21.50, 25.00)	23.50 (21.75, 25.00)	23.50 (22.00, 25.00)	21.00 (20.00, 22.00)	22.00 (20.53, 24.00)	23.00 (21.25, 24.38)	23.50 (21.81, 25.50)	
Unknown	9 (14%)	29 (4.6%)	0 (0%)	2 (3.4%)	0 (0%)	6 (19%)	30 (44%)	0 (0%)	
Delivery Method									0.040
Viginal delivery	16 (25%)	176 (28%)	15 (28%)	18 (31%)	14 (35%)	10 (31%)	17 (25%)	14 (37%)	
Cesarean section	48 (74%)	453 (72%)	39 (72%)	41 (69%)	26 (65%)	22 (69%)	49 (72%)	24 (63%)	
Unknown	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (2.9%)	0 (0%)	
Prenatal Corticosteriods									<0.001
No	5 (7.7%)	85 (14%)	6 (11%)	12 (20%)	3 (7.5%)	4 (13%)	5 (7.4%)	4 (11%)	
Yes	56 (86%)	543 (86%)	46 (85%)	46 (78%)	37 (93%)	26 (81%)	41 (60%)	33 (87%)	
Unknown	4 (6.2%)	1 (0.2%)	2 (3.7%)	1 (1.7%)	0 (0%)	2 (6.3%)	22 (32%)	1 (2.6%)	
Complete Prenatal Steriods									<0.001
No	15 (23%)	110 (17%)	6 (11%)	18 (31%)	10 (25%)	8 (25%)	17 (25%)	7 (18%)	
Yes	34 (52%)	414 (66%)	40 (74%)	24 (41%)	27 (68%)	15 (47%)	26 (38%)	26 (68%)	
Unknown	16 (25%)	105 (17%)	8 (15%)	17 (29%)	3 (7.5%)	9 (28%)	25 (37%)	5 (13%)	
Maternal Chorioamniontis									<0.001
No	18 (28%)	524 (83%)	28 (52%)	50 (85%)	25 (63%)	28 (88%)	63 (93%)	31 (82%)	
Yes	17 (26%)	105 (17%)	4 (7.4%)	8 (14%)	14 (35%)	3 (9.4%)	5 (7.4%)	2 (5.3%)	
Unknown	30 (46%)	0 (0%)	22 (41%)	1 (1.7%)	1 (2.5%)	1 (3.1%)	0 (0%)	5 (13%)	
Gender									0.7
Female	26 (40%)	248 (39%)	21 (39%)	28 (47%)	17 (43%)	16 (50%)	28 (41%)	20 (53%)	
Male	38 (58%)	379 (60%)	32 (59%)	31 (53%)	23 (58%)	16 (50%)	40 (59%)	18 (47%)	
Unknown	1 (1.5%)	2 (0.3%)	1 (1.9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Small for Gestational Age									0.004
Not SGA	38 (58%)	502 (80%)	40 (74%)	53 (90%)	32 (80%)	24 (75%)	51 (75%)	31 (82%)	
SGA	26 (40%)	117 (19%)	11 (20%)	5 (8.5%)	8 (20%)	8 (25%)	17 (25%)	7 (18%)	
Unknown	1 (1.5%)	10 (1.6%)	3 (5.6%)	1 (1.7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	

[1] n (%)

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

Figure 1: Summary table of the characteristics stratified by center.

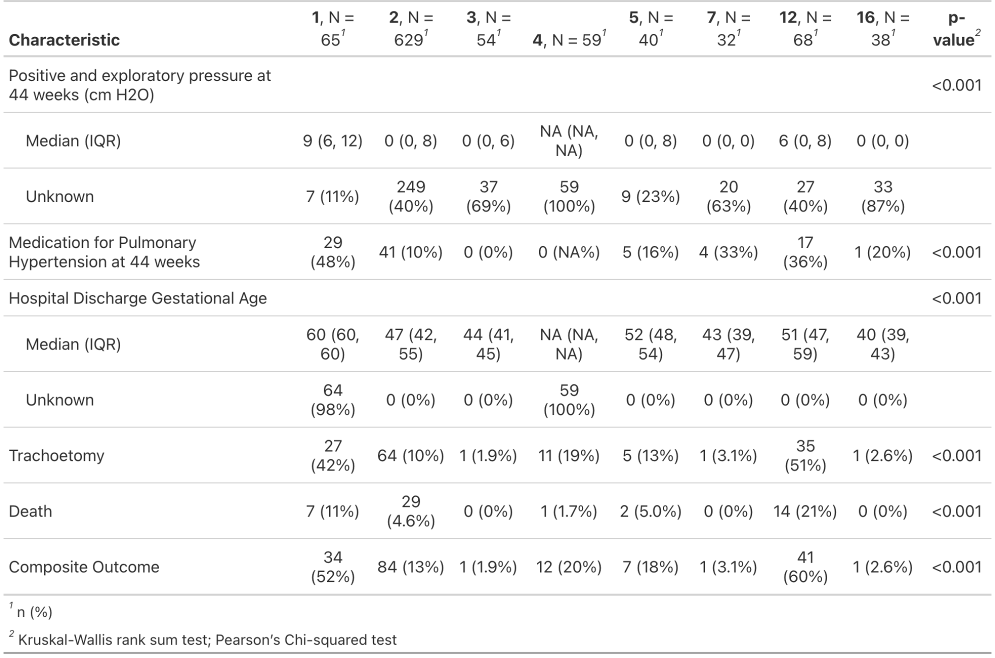


Characteristic	**1**, N = 65[1]	**2**, N = 629[1]	**3**, N = 54[1]	**4**, N = 59[1]	**5**, N = 40[1]	**7**, N = 32[1]	**12**, N = 68[1]	**16**, N = 38[1]	**p-value**[2]
Surfactant used in the first 72 hrs									<0.001
No	4 (6.2%)	70 (11%)	2 (3.7%)	4 (6.8%)	2 (5.0%)	3 (9.4%)	9 (13%)	7 (18%)	
Yes	36 (55%)	265 (42%)	51 (94%)	11 (19%)	33 (83%)	3 (9.4%)	47 (69%)	9 (24%)	
Unknown	25 (38%)	294 (47%)	1 (1.9%)	44 (75%)	5 (13%)	26 (81%)	12 (18%)	22 (58%)	
Weight at 36 weeks (g)									0.011
Median (IQR)	2,100 (1,843, 2,354)	2,150 (1,865, 2,405)	2,130 (1,933, 2,445)	2,100 (1,860, 2,400)	1,943 (1,723, 2,138)	2,200 (1,925, 2,420)	2,020 (1,785, 2,180)	2,273 (2,069, 2,472)	
Unknown	17 (26%)	36 (5.7%)	3 (5.6%)	6 (10%)	0 (0%)	1 (3.1%)	29 (43%)	0 (0%)	
Ventilation Support at 36 weeks									<0.001
No respiratory support or suppl. oxygen	8 (12%)	49 (7.8%)	5 (9.3%)	8 (14%)	0 (0%)	22 (69%)	1 (1.5%)	22 (58%)	
Non-invasive positive pressue	22 (34%)	425 (68%)	34 (63%)	34 (58%)	31 (78%)	8 (25%)	16 (24%)	14 (37%)	
Invasive positive pressure	34 (52%)	146 (23%)	14 (26%)	17 (29%)	9 (23%)	2 (6.3%)	32 (47%)	2 (5.3%)	
Unknown	1 (1.5%)	9 (1.4%)	1 (1.9%)	0 (0%)	0 (0%)	0 (0%)	19 (28%)	0 (0%)	
Fraction of Inspired Oxygen at 36 weeks									<0.001
Median (IQR)	0.35 (0.30, 0.49)	0.27 (0.23, 0.35)	0.30 (0.25, 0.35)	0.40 (0.30, 0.50)	0.33 (0.26, 0.43)	0.35 (0.32, 0.38)	0.35 (0.28, 0.45)	0.35 (0.27, 0.39)	
Unknown	18 (28%)	36 (5.7%)	2 (3.7%)	3 (5.1%)	0 (0%)	1 (3.1%)	29 (43%)	0 (0%)	
Peak Inspiratory Pressure at 36 weeks (cm H2O)									<0.001
Median (IQR)	2 (0, 14)	0 (0, 0)	0 (0, 15)	4 (0, 9)	0 (0, 9)	0 (0, 0)	12 (0, 15)	0 (0, 0)	
Unknown	20 (31%)	39 (6.2%)	5 (9.3%)	15 (25%)	13 (33%)	1 (3.1%)	32 (47%)	0 (0%)	
Positive and exploratory pressure at 36 weeks (cm H2O)									<0.001
Median (IQR)	8 (6, 9)	7 (6, 8)	8 (7, 10)	6 (6, 7)	9 (8, 10)	0 (0, 5)	6 (6, 7)	0 (0, 8)	
Unknown	24 (37%)	41 (6.5%)	9 (17%)	6 (10%)	0 (0%)	1 (3.1%)	34 (50%)	0 (0%)	
Medication for Pulmonary Hypertension at 36 weeks	13 (20%)	25 (4.0%)	3 (5.7%)	10 (17%)	3 (7.5%)	2 (6.3%)	4 (8.2%)	4 (11%)	<0.001
Weight at 44 weeks (g)									0.007
Median (IQR)	3,680 (3,338, 4,200)	3,768 (3,376, 4,140)	3,820 (3,360, 4,220)	NA (NA, NA)	3,372 (3,155, 3,998)	3,860 (3,375, 4,465)	3,270 (2,903, 3,815)	2,950 (2,468, 4,110)	
Unknown	7 (11%)	251 (40%)	37 (69%)	59 (100%)	9 (23%)	21 (66%)	25 (37%)	33 (87%)	
Ventilation Support at 44 weeks									<0.001
No respiratory support or suppl. oxygen	11 (17%)	198 (31%)	12 (22%)	0 (0%)	19 (48%)	10 (31%)	12 (18%)	5 (13%)	
Non-invasive positive pressue	18 (28%)	97 (15%)	6 (11%)	0 (0%)	9 (23%)	0 (0%)	13 (19%)	0 (0%)	
Invasive positive pressure	32 (49%)	96 (15%)	0 (0%)	0 (0%)	3 (7.5%)	2 (6.3%)	22 (32%)	0 (0%)	
Unknown	4 (6.2%)	238 (38%)	36 (67%)	59 (100%)	9 (23%)	20 (63%)	21 (31%)	33 (87%)	
Fraction of Inspired Oxygen at 44 weeks									0.010
Median (IQR)	0.31 (0.25, 0.42)	0.28 (0.26, 0.35)	0.25 (0.23, 0.28)	NA (NA, NA)	0.27 (0.24, 0.33)	0.31 (0.25, 0.47)	0.31 (0.25, 0.51)	0.27 (0.24, 0.29)	
Unknown	9 (14%)	252 (40%)	37 (69%)	59 (100%)	9 (23%)	21 (66%)	24 (35%)	33 (87%)	
Peak Inspiratory Pressure at 44 weeks (cm H2O)									<0.001
Median (IQR)	10 (0, 17)	0 (0, 2)	0 (0, 0)	NA (NA, NA)	0 (0, 0)	0 (0, 0)	0 (0, 17)	0 (0, 0)	
Unknown	7 (11%)	247 (39%)	36 (67%)	59 (100%)	14 (35%)	22 (69%)	25 (37%)	33 (87%)	

[1] n (%)

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

Figure 2: (Continued) Summary table of the characteristics stratified by center.



Characteristic	**1**, N = 65[1]	**2**, N = 629[1]	**3**, N = 54[1]	**4**, N = 59[1]	**5**, N = 40[1]	**7**, N = 32[1]	**12**, N = 68[1]	**16**, N = 38[1]	**p-value**[2]
Positive and exploratory pressure at 44 weeks (cm H2O)									<0.001
Median (IQR)	9 (6, 12)	0 (0, 8)	0 (0, 6)	NA (NA, NA)	0 (0, 8)	0 (0, 0)	6 (0, 8)	0 (0, 0)	
Unknown	7 (11%)	249 (40%)	37 (69%)	59 (100%)	9 (23%)	20 (63%)	27 (40%)	33 (87%)	
Medication for Pulmonary Hypertension at 44 weeks	29 (48%)	41 (10%)	0 (0%)	0 (NA%)	5 (16%)	4 (33%)	17 (36%)	1 (20%)	<0.001
Hospital Discharge Gestational Age									<0.001
Median (IQR)	60 (60, 60)	47 (42, 55)	44 (41, 45)	NA (NA, NA)	52 (48, 54)	43 (39, 47)	51 (47, 59)	40 (39, 43)	
Unknown	64 (98%)	0 (0%)	0 (0%)	59 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Trachoetomy	27 (42%)	64 (10%)	1 (1.9%)	11 (19%)	5 (13%)	1 (3.1%)	35 (51%)	1 (2.6%)	<0.001
Death	7 (11%)	29 (4.6%)	0 (0%)	1 (1.7%)	2 (5.0%)	0 (0%)	14 (21%)	0 (0%)	<0.001
Composite Outcome	34 (52%)	84 (13%)	1 (1.9%)	12 (20%)	7 (18%)	1 (3.1%)	41 (60%)	1 (2.6%)	<0.001

[1] n (%)

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

Figure 3: (Continued) Summary table of the characteristics stratified by center.

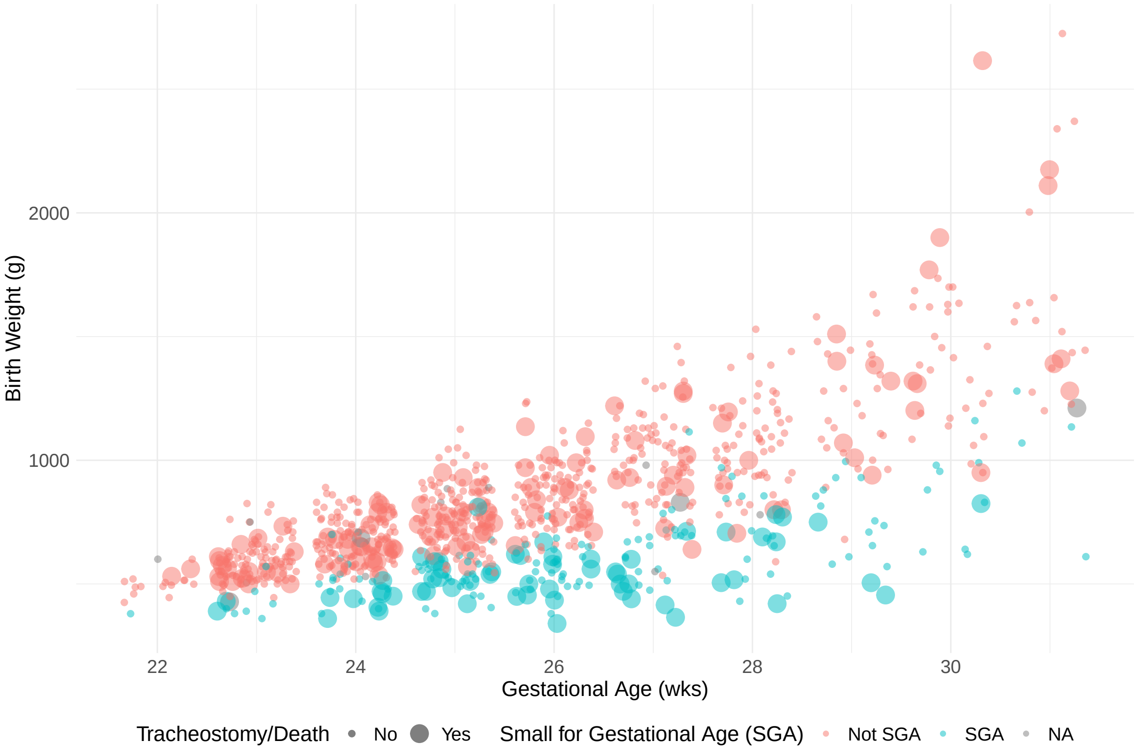


Figure 4: Association between birth weight and gestational age at birth. Points are colored by SGA and scaled by the composite outcome.

one. For example, the maximum difference in intercepts is about 0.7 in the first lasso model, but the difference between center 2 and center 12 is more than 1 in multilevel lasso.

The general trend regarding intercepts and coefficient estimates across models is similar for 44-week models. Some variables like maternal chorioamniontis and positive and exploratory pressure at 36 weeks were selected out for 44-week models. This could be due to the collinearity between 36-week and 44-week measurements or the difference in the sub-populations at different timing.

On the validation data, we see the multilevel lasso model achives the highest AUC (0.810; 0.855), which are higher than the lasso model with center at 36 weeks (0.904) but lower than it at 44 weeks (0.858). They are both higher than the one without center (0.781; 0.849) (Table 3, 4). The lasso model with center has a highest F-score among three models for each timing. At their optimal thresholds, we see multilevel lasso has the highest specificity and sensitivity at 36 weeks, while the lasso with center has the highest specificity at 44 weeks. The multilevel lasso has the highest sensitivity at 44 weeks. We see all models improve at 44 weeks compared to 36 weeks as expected. This is could be due to we include more information with more predictors, or it could be due to the different between the two sub-populations at different timing.

Table 1: Coefficient estimates from 3 models at 36 weeks. The intercepts for Lasso models are calculated by combining the overall intercept and the individual center intercept, while the model without center doesn't have individual intercepts. Coefficients are rounded to three digits for display.

	Lasso (with center)	Lasso (without center)	Multilevel Lasso
Center 1	-2.14	-2.021	-1.468
Center 2	-2.435	-2.021	-2.215
Center 3	-2.462	-2.021	-1.982
Center 4	-2.14	-2.021	-1.693
Center 5	-2.14	-2.021	-1.747
Center 7	-2.14	-2.021	-1.777
Center 12	-1.794	-2.021	-1.175
Center 16	-2.175	-2.021	-1.815
Not Hispanic or Latino	0.244	0.222	0.233
Birth Weight	0.005	0.005	0.003
Gestational Age			0.014
Birth length			0.009
Birth Head Circumference	0.219	0.23	0.165
Delivery Method: Cesarean section	0.063	0.093	0.075
Prenatal Corticosteriods: Yes	0.302	0.329	0.23
Complete Prenatal Steriods: Yes	0.025	0.012	0.041
Maternal Chorioamniontis: Yes	0.024	0.021	0.022
Gender: Male			
SGA	0.102	0.109	0.068
Surfactant: Yes			
Surfactant: No			-0.009
Weight at 36 weeks	-0.044	-0.056	-0.053
Ventilation Support at 36 weeks: Non-invasive	-0.002		-0.097
Ventilation Support at 36 weeks: Invasive	0.441	0.731	0.432
Inspired Oxygen at 36 weeks	0.682	0.746	0.631
Peak Inspiratory Pressure at 36 weeks	0.349	0.062	0.318
Positive and exploratory pressure at 36 weeks	0.199	0.118	0.148
Medication for Pulmonary Hypertension at 36 weeks	0.065	0.095	0.057

Table 2: Coefficient estimates from 3 models at 44 weeks.

	Lasso (with center)	Lasso (without center)	Multilevel Lasso
Center 1	-2.065	-1.946	-1.612
Center 2	-2.088	-1.946	-1.991
Center 3	-2.235	-1.946	-1.867
Center 4	-2.065	-1.946	-1.769
Center 5	-2.065	-1.946	-1.745
Center 7	-2.123	-1.946	-1.809
Center 12	-1.678	-1.946	-1.404
Center 16	-2.238	-1.946	-1.868
Not Hispanic or Latino	0.047	0.031	0.103
Birth Weight			-0.004
Gestational Age			0.046
Birth length			-0.014
Birth Head Circumference	0.036	0.014	0.013
Delivery Method: Cesarean section	0.134	0.114	0.138
Prenatal Corticosteriods: Yes	0.312	0.257	0.254
Complete Prenatal Steriods: Yes	0.013		0.012
Maternal Chorioamniontis: Yes			
Gender: Male			
SGA	0.137	0.072	0.111
Surfactant: Yes			0.004
Surfactant: No			0.016
Weight at 36 weeks	-0.065	-0.034	-0.06
Ventilation Support at 36 weeks: Non-invasive			-0.166
Ventilation Support at 36 weeks: Invasive	0.68	0.772	0.489
Inspired Oxygen at 36 weeks	0.345	0.379	0.326
Peak Inspiratory Pressure at 36 weeks	0.032		0.135
Positive and exploratory pressure at 36 weeks			0.067
Medication for Pulmonary Hypertension at 36 weeks	0.106	0.044	0.127
Weight at 44 weeks	-0.178	-0.278	-0.235
Ventilation Support at 44 weeks: Non-invasive	-0.089		-0.037
Ventilation Support at 44 weeks: Invasive	0.24	0.274	0.235
Inspired Oxygen at 44 weeks	0.08	0.086	0.131
Peak Inspiratory Pressure at 44 weeks			0.03
Positive and exploratory pressure at 44 weeks	0.396	0.323	0.305
Medication for Pulmonary Hypertension at 44 weeks	0.311	0.341	0.275

In ROC curves, we see multilevel Lasso has the best performance at high specificity compared to the other two at 36 weeks, and it becomes comparable with the lasso model with center at 44 weeks. The lasso model without center tends to be the worst (Figure 5, 6).

For calibration, we chose to examine the multilevel lasso regression model since it performs overall the best. At 36 weeks, we see the model is poorly calibrated as many confidence intervals doesn't cover the identity line. The calibration is improved for 44 weeks as the identity line is covered by more bars. Unexpectedly, we see the observed proportion doesn't increase all the time when the expected proportion increases, specially around expected proportion = 0.75. In addition, the overall expected proportion from the model tends to overestimate the observed proportion, resulting in relatively higher sensitivity but low specificity.

Discussion

This analysis of premature infants with BPD illustrates the successful development of prediction model on tracheotomy/death composite outcome. Our multilevel lasso model result in a high prediction accuracy, and it not only obtains different baseline estimates for different centers but also account for the random errors across centers. Our model fitting setting illustrates three different scenarios: pooling all observations by not incorporating centers, partial pooling by using centers as predictors, and partial pooling by using centers as random effects. To obtain a higher accuracy, it is necessary to incorporate center information. In the exploratory analysis, center with similar baseline covariates can have significantly different outcomes, like center 1 and center 5. This baseline difference and difference in sample sizes could be related to the hierarchy of hospitals. For example, a tertiary referral hospital may have more advanced instruments and experiences for neonatal care while it may receive more premature infants with serious complications. Multilevel modeling can potentially provide a more precise accurate for baseline level for each center by estimating between-center and within-center variability. Although no variables are selected out by the multilevel modeling, we can set low coefficients to be 0 to get a simpler model but with comparable performance.

Also, our estimates for the 36-weeks and 44-week measurements are consistent with the previous literature. We see the coefficients for invasive ventilation support and inspired oxygen are positive and large, and these two variables are associated with the lung underdevelopment. The predictor for prenatal corticosteriods also has a high positive coefficient, which is consistent with the beneficial effects of steroids in infant development (*Pregnancy and birth* 2018). The coefficient of non-invasive ventilation support is negative for all three models, implying that infants with non-invasive ventilation unlikely need tracheotomy at discharge. The predictor hospital discharge gestational age has the highest coefficient value, meaning the longer the stay is, the worse the health outcome the infatns have. Surprisingly, gestational age, birth length, and birth weight are not included in the two lasso models, and their coefficients are close to 0 in multilevel lasso model. This can be due to that other variables of clinical decisions are associated these measurements at birth and thus absorb their effects, which is similar to the collinearity we observed in exploratory analysis.

Our study has several limitations. First, the missing proportion varies for different variables at different centers. Our multiple imputation based on the full data may result in biased estimates for these variables. Second, our performance metrics account for the imbalance in the outcome but not account for the imbalance in the center size. Our model may yield good performance overall but incorrect predictions in some small centers. Third, our multilevel lasso and the lasso model with center both require sufficient existing observations for one center. They can't do prediction on a new data belonging to a center that doesn't exist in the train set. Finally, our models have poor calibration, which seems to be the major problem.

The multilevel predictive models can be extended. For example, we can consider random slopes for predictors related to clinical decisions - different centers may have different general guideline when giving some medication and doing some tests. This is also associated with the missing proportion from the exploratory analysis as it is possible that some measurements won't be conducted in some centers usually.

Conclusion

This paper develops three prediction models on tracheotomy or death for premature babies with different approaches, and multilevel lasso model has the best performance if the center information can be obtained with a relatively large sample size. This approach accounts for the heterogeneity of clinical setting of different hospitals and also measurements at different timing. The clinical application of these models is to estimate whether premature infants require additional care and resources if they likely result in tracheotomy or death in the future. Our expectation is that the clinical care can be individualized for children with a lower mortality rate.

Table 3: Model performance on the evaluation data set at 36 weeks. The values are obtained at the optimal threshold of each model determined by Youden's J statistics.

	Lasso (with center)	Lasso (without center)	Multilevel Lasso
Threshold	0.104	0.138	0.135
Sensitivity	0.769	0.742	0.780
Specificity	0.772	0.773	0.762
AUC	0.793	0.781	0.810
F-score	0.551	0.537	0.548

Table 4: Model performance on the evaluation data set at 44 weeks. The values are obtained at the optimal threshold of each model determined by Youden's J statistics.

	Lasso (with center)	Lasso (without center)	Multilevel Lasso
Threshold	0.191	0.211	0.212
Sensitivity	0.813	0.806	0.819
Specificity	0.809	0.770	0.771
AUC	0.858	0.849	0.855
F-score	0.555	0.512	0.519

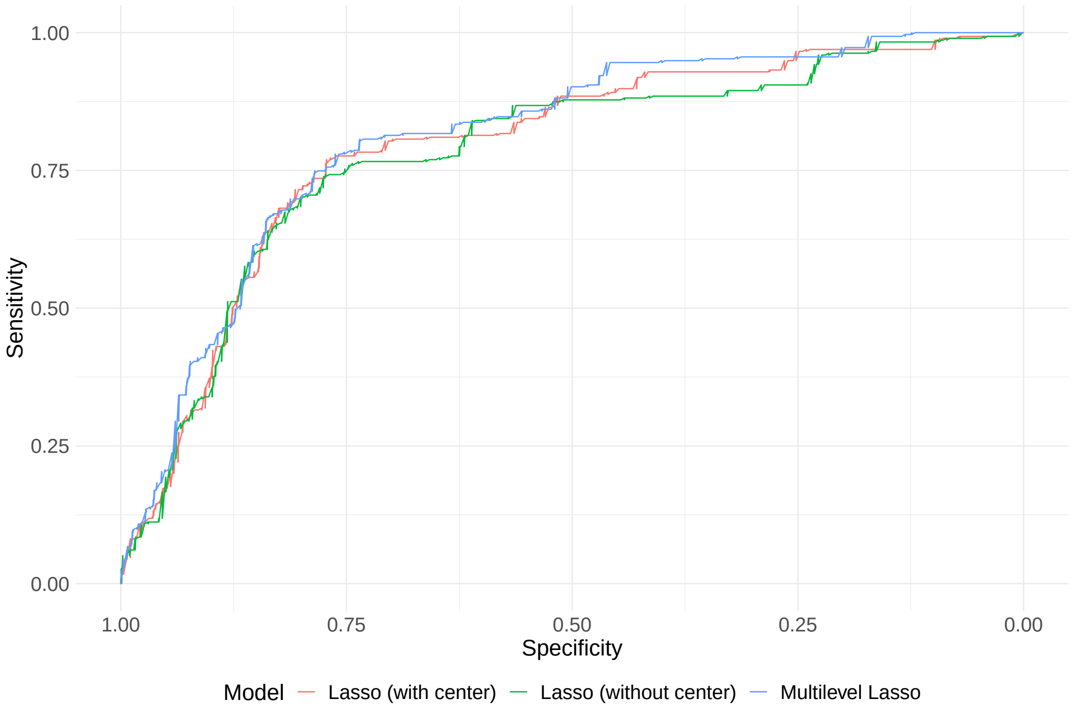


Figure 5: ROC curves for 3 models on the evaluation set at 36 weeks.

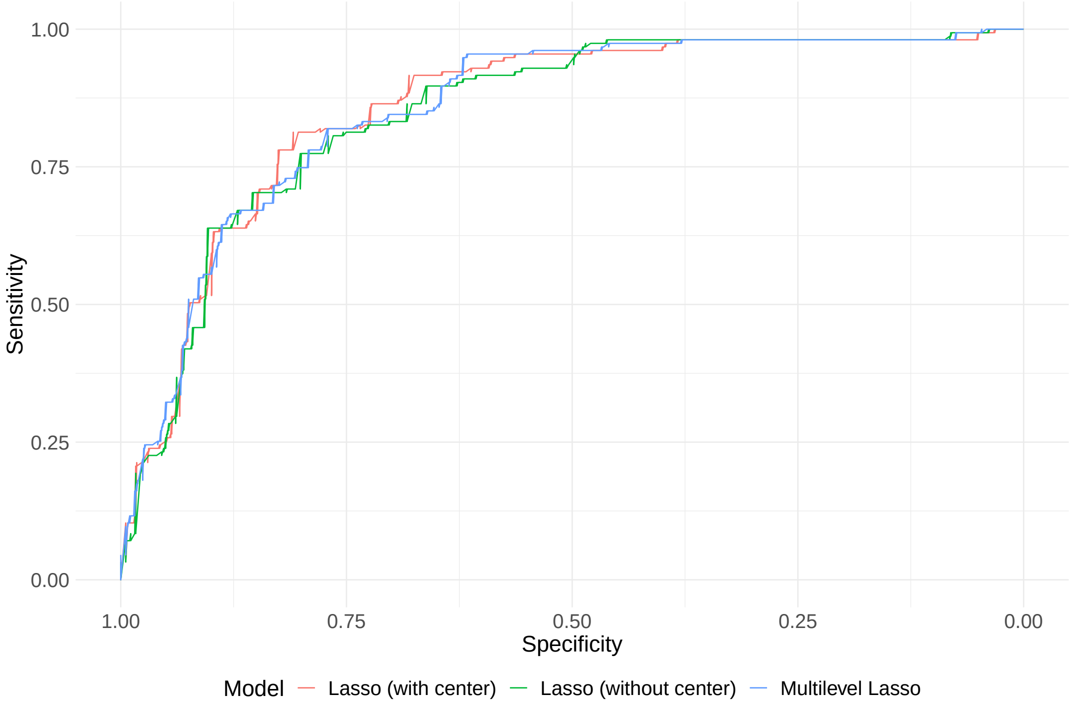


Figure 6: ROC curves for 3 models on the evaluation set at 44 weeks.

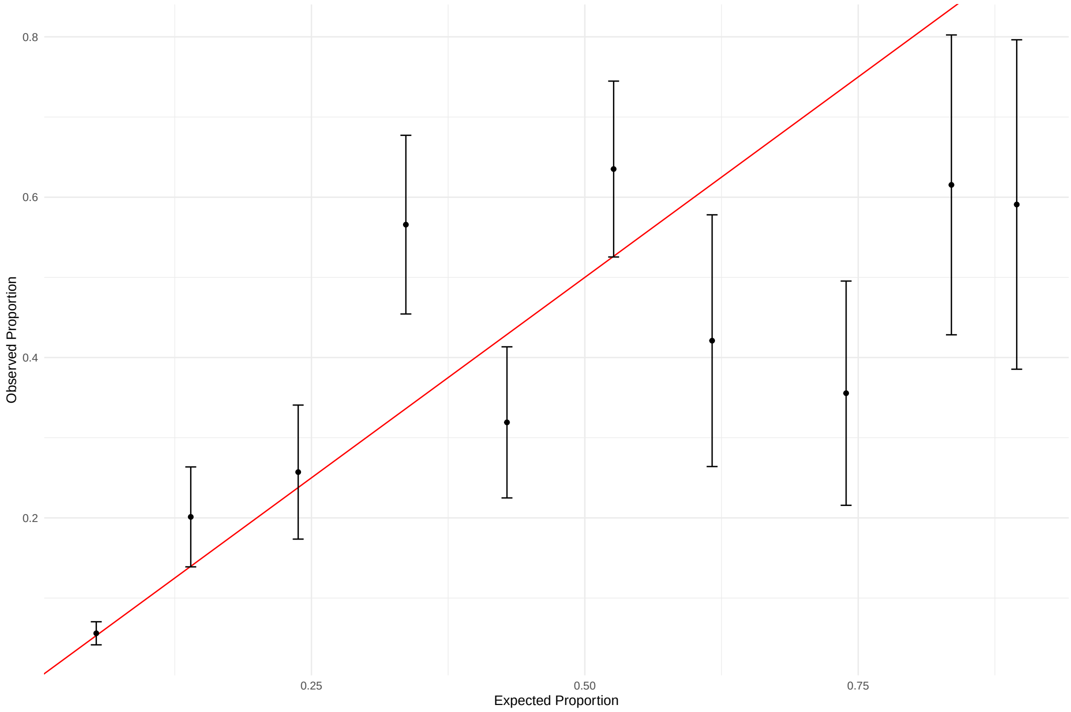


Figure 7: Calibration plot of the Multilevel Lasso regression at 36 weeks.

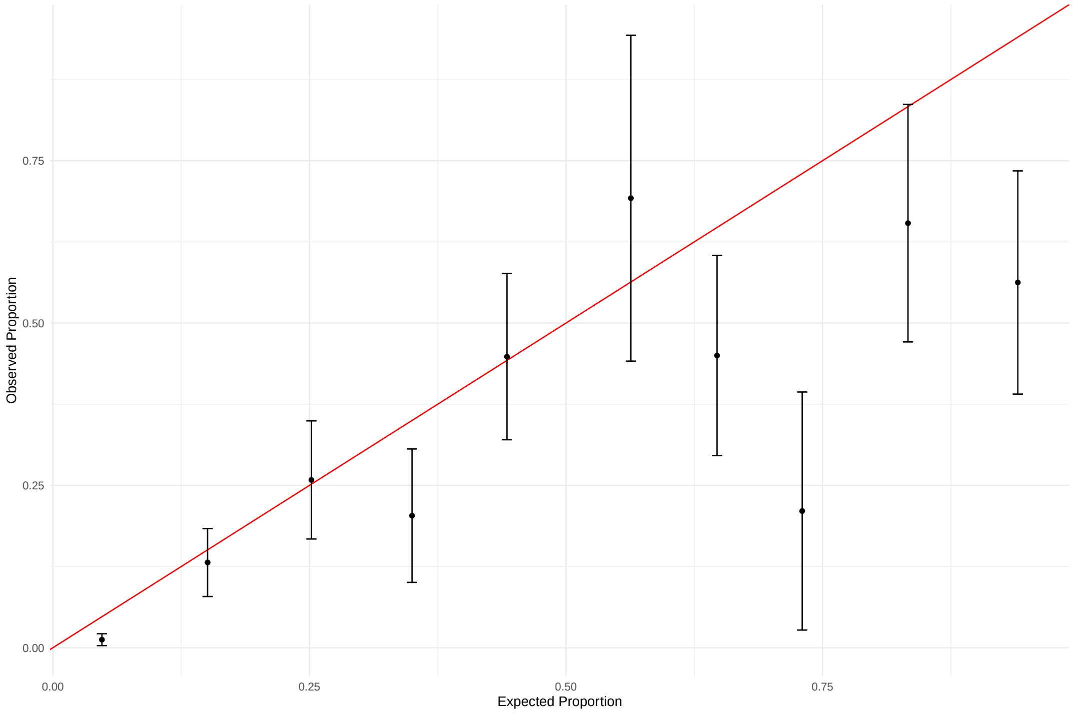


Figure 8: Calibration plot of the Multilevel Lasso regression at 44 weeks.

References

Friedman, J., Hastie, T., and Tibshirani, R. (2010), “Regularization paths for generalized linear models via coordinate descent,” *J. Stat. Softw.*, Foundation for Open Access Statistic, 33, 1–22.

Groll, A. (2023), *glmmLasso: Variable selection for generalized linear mixed models byL1-penalized estimation.*

He, Y. (2010), “Missing data analysis using multiple imputation: Getting to the heart of the matter,” *Circ. Cardiovasc. Qual. Outcomes*, Ovid Technologies (Wolters Kluwer Health), 3, 98–105.

Jain, V. G., Willis, K. A., Jobe, A., and Ambalavanan, N. (2022), “Chorioamnionitis and neonatal outcomes,” *Pediatr. Res.*, Springer Science; Business Media LLC, 91, 289–296.

Kalikkot Thekkeveedu, R., Guaman, M. C., and Shivanna, B. (2017), “Bronchopulmonary dysplasia: A review of pathogenesis and pathophysiology,” *Respir. Med.*, 132, 170–177.

Murthy, K., Savani, R. C., Lagatta, J. M., Zaniletti, I., Wadhawan, R., Truog, W., Grover, T. R., Zhang, H., Asselin, J. M., Durand, D. J., Short, B. L., Pallotto, E. K., Padula, M. A., Dykes, F. D., Reber, K. M., and Evans, J. R. (2014), “Predicting death or tracheostomy placement in infants with severe bronchopulmonary dysplasia,” *J. Perinatol.*, Springer Science; Business Media LLC, 34, 543–548.

Paul, A. (2023), “PHP 2550: Practical data analysis,” PHP 2550.

Pregnancy and birth: Before preterm birth: What do steroids do? (2018), Institute for Quality; Efficiency in Health Care (IQWiG).

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., and Schisterman, E. F. (2008), “Youden index and optimal cut-point estimated from observations affected by a lower limit of detection,” *Biom. J.*, 50, 419–430.

Unal, S., Bilgin, L. K., Gonulal, D., and Akcan, F. A. (2015), “Optimal time of tracheotomy in infants: Still a dilemma,” *Glob. Pediatr. Health*, SAGE Publications, 2, 2333794X15569300.

van Buuren, S., and Groothuis-Oudshoorn, K. (2011), “mice: Multivariate imputation by chained equations in r,” *Journal of Statistical Software*, 45, 1–67. https://doi.org/10.18637/jss.v045.i03.