Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Postoperative neonatal mortality prediction using superlearning

Jennifer N. Cooper, PhD,[a,*] *Peter C. Minneci, MD, MHSc,*[a,b] *and Katherine J. Deans, MD, MHSc*[a,b]

[a] *Center for Surgical Outcomes Research and Center for Innovation in Pediatric Practice, The Research Institute at Nationwide Children's Hospital, Columbus, Ohio*

[b] *Department of Surgery, Nationwide Children's Hospital, Columbus, Ohio*

ARTICLE INFO

Article history:
Received 17 January 2017
Received in revised form
7 August 2017
Accepted 1 September 2017

Keywords:
Superlearning
Prediction
Neonates
Postoperative mortality

ABSTRACT

Background: The variable risks associated with neonatal surgery present a challenge to accurate mortality prediction. We aimed to apply superlearning, an ensemble machine learning method, to the prediction of 30-day neonatal postoperative mortality.

Materials and methods: We included neonates in the 2012-2014 National Surgical Quality Improvement Program Pediatric. Patients treated in 2012-13 were used in model development ($n = 6499$), and patients treated in 2014 formed the validation sample ($n = 3552$). Our superlearner algorithm included 14 regression and machine learning algorithms and included all preoperative patient demographic and clinical characteristics, including indicator variables for surgical procedures. Performance was evaluated using mean squared error and measures of discrimination and calibration.

Results: The superlearner out-performed all individual algorithms with regard to cross-validated mean squared error. It showed excellent discrimination, with an area under the receiver-operating characteristic curve of 0.91 in development and 0.87 in validation. The superlearner showed good calibration in development but not in validation (Cox calibration test $P = 0.06$ and $P < 0.001$, respectively). Performance was improved when the superlearner was fit using only variables strongly associated with mortality in bivariate analysis (area under the receiver-operating characteristic curve 0.89, calibration test $P = 0.63$ in validation).

Conclusions: Superlearning provided improved or equivalent performance compared with individual regression and machine learning algorithms for predicting neonatal surgical mortality. This method should be considered for prediction in large data sets whenever complex mechanisms make parametric modeling assumptions unrealistic.

Introduction

Neonates undergoing surgical procedures are at greater risk of postoperative morbidity and mortality than older children and adults.[1,2] Neonatal surgery patients vary widely in their indications for surgery and in their comorbidities. The variable risks associated with neonatal procedures and diseases, many of which are rare, present a challenge to the accurate prediction of mortality and morbidity in individual patients. This patient heterogeneity, combined with the infrequency with which individual tertiary children's hospitals perform most types of neonatal surgical procedures, necessitate the use of large

* *Corresponding author*. Center for Surgical Outcomes Research and Center for Innovation in Pediatric Practice, The Research Institute at Nationwide Children's Hospital, 700 Children's Drive, FB Suite 3A.3, Columbus, OH, 43205. Tel.: +1 614-355-4526; fax: +1 614-722-3544.
E-mail address: jennifer.cooper@nationwidechildrens.org (J.N. Cooper).

0022-4804/$ – see front matter
https://doi.org/10.1016/j.jss.2017.09.002

multiinstitutional databases for the development of prediction models for neonatal postoperative mortality. Such models can be used to better inform families of their infant's risk of death after undergoing a surgical procedure.

Several studies have attempted to develop clinical prediction models for postoperative mortality in neonates undergoing major noncardiac surgery.[1,3,4] One such study used administrative hospital discharge databases whereas the others used the American College of Surgeons' National Surgical Quality Improvement Program Pediatric (ACS-NSQIP-P) database, which contains standardized, validated preoperative clinical data, and procedural data. All three studies developed logistic regression models that achieved good discrimination, suggesting that these models would perform adequately for risk adjustment across groups. Logistic regression, however, can be limited for achieving accurate patient-level outcome prediction because it imposes stringent, parametric constraints on the relationship between predictors and the probability of an outcome. For instance, main-terms logistic regression relies on the assumption that relationships between a prespecified transformation of the outcome and its predictors are both linear and additive. Biased predictions result when these assumptions are untrue. Given the complex and variable processes underlying mortality in neonatal surgery patients, such assumptions may be unrealistic. In contrast to parametric regression models, superlearning is an ensemble machine learning method for selecting, via cross-validation, the optimal prediction algorithm among all weighted combinations of a set of candidate algorithms, which can include both parametric regression models and data-adaptive algorithms.[5] The aim of this study was to develop and validate a clinical prediction model, using the flexible superlearning approach, for 30-day postoperative mortality in neonates. Because previous studies have reported preterm neonates to be at a much higher risk of postsurgical mortality than term neonates, we also aimed to examine the performance of our derived model in preterm neonates specifically and examine predicted mortality rates by gestational age.

Materials and methods

Data source and study population

The conduct of this study was approved by the Nationwide Children's Hospital Institutional Review Board with a waiver of informed consent. This study used the 2012-14 Participant Use Files (PUFs) of the ACS-NSQIP-P. During this period, between 50 and 59 U.S. and Canadian children's hospitals participated in ACS-NSQIP-P each year. The available data consisted of demographic, preoperative clinical and procedural variables and postoperative adverse events, including mortality, in the 30 days following surgery, as previously described.[6-8] ACS-NSQIP-P is a multispecialty program with cases sampled from pediatric general/thoracic surgery, otolaryngology, orthopedic surgery, urology, neurosurgery, and plastic surgery. The program provides peer-reviewed, risk-adjusted 30-day postoperative outcomes to participating institutions, for the purposes of benchmarking and quality improvement. Included cases are selected based on current procedural terminology (CPT) codes using 8-day cycle-based systematic sampling of 35 procedures per cycle. Data are collected from patient medical records and directly from patients and their families.[6-8] All surgical cases with CPT codes on the ACS-NSQIP-P CPT code inclusion list are included, with the exception of procedures in patients over 18 years of age, procedures performed due to complications of previous procedures, trauma cases, and transplant cases. In this study, only neonatal patients were included. Consistent with the ACS-NSQIP-P definition, neonates were defined as infants who were either (1) born at term (greater than 37 weeks gestation) and were less than 29 days old at the time of surgery or (2) born preterm (less than 37 weeks gestation) and had a gestational age at surgery of less than 51 weeks.

Preoperative, procedural, and outcome variables

For this study, all available preoperative variables were included in analyses, with the exception of comorbidities that were present in fewer than 10 neonates and variables that had missing values in more than 30% of the neonates. ACS-NSQIP-P collects several variables specifically in neonates, namely gestational ages at birth and at surgery, inborn status, APGAR scores at 1 and 5 minutes after birth, mode of delivery, birth weight, height, and head circumference. As birth height and head circumference were missing in more than 30% of neonates, these variables were not included in analyses. Several other variables were also missing in large numbers of patients. However, as many of these variables were preoperative lab values, we chose not to use multiple imputations to fill in missing values as we doubted that the other variables included in the data set would enable accurate estimation of probable values for the missing data. Instead, all variables with missing values were categorized, at their quartiles if continuous, and missing values simply formed another category. In total, sixty-eight variables were considered as predictors of mortality. However, to consider procedures as predictors, a binary variable was created for each CPT code that was present in at least 10 cases. This was done separately for primary procedures and for the other procedures that were performed concurrently in some patients. Similarly, binary variables were created for each congenital anomaly (as defined by International Classification of Diseases 9th Revision, Clinical Modification diagnosis codes) present in at least 10 cases. After the creation of these binary indicator variables, a total of 284 preoperative characteristics and procedure variables were available for outcome prediction model building. The outcome of interest was mortality within 30 days after surgery, whether in the hospital or after discharge.

Development and validation of the superlearner

The superlearner algorithm (SL) was used for prediction model building in this study. The SL was proposed by van der Laan *et al.*[5] as a method for selecting the optimal regression algorithm among all weighted combinations of a set of candidate algorithms. The analyst provides a collection of prediction algorithms and specifies a loss function, which in this study was the squared difference between observed and predicted outcomes. The SL then uses V-fold cross-validation

to estimate the mean squared prediction error (CV-MSE) of each algorithm on data not used in building the model, and then selects the weighted linear convex combination of algorithms that provides the smallest squared prediction error. Based on the theory behind superlearning, to optimize performance, the library of candidate algorithms should include as many algorithms as possible. This study used 14 algorithms to limit the computational resources required for the analysis. We used 10-fold cross-validation and included the following candidate algorithms: 1) a pruned classification and regression tree (CART) with a complexity parameter of 0.01 2) a pruned CART with a complexity parameter of 0.001[9] 3) a logistic regression model fit using forward variable selection 4) a logistic regression model fit using stepwise variable selection 5) a logistic regression model fit via lasso-penalized maximum likelihood 6) a logistic regression model fit via penalized maximum likelihood with an elastic-net mixing parameter of 0.5 7) a logistic regression model fit via penalized maximum likelihood with an elastic-net mixing parameter of 0.2[10] 8) boosted CART with an interaction depth of 2 9) boosted CART with an interaction depth of 3[11,12] 10) a random forest with 94 predictors randomly sampled as candidates at each split and a minimum terminal node size of 1 11) a random forest with 50 predictors randomly sampled as candidates at each split and a minimum terminal node size of 1 12) a random forest with 25 predictors randomly sampled as candidates at each split and a minimum terminal node size of 1 13) a random forest with 50 predictors randomly sampled as candidates at each split and a minimum terminal node size of 5, and 14) a random forest with 25 predictors randomly sampled as candidates at each split and a minimum terminal node size of 5. To assess the true performance of the SL, an additional layer of 10-fold cross-validation was performed and the performance measures described below were assessed.

The discrimination of the SL and each of its candidate algorithms was assessed with the cross-validated AUROC (CV-AUROC) and graphically depicted using receiver-operating characteristic curves.[13] Model calibration was assessed with the Cox calibration test.[14,15] Because of its many limitations, including inadequate performance in large samples, the more conventional Hosmer–Lemeshow statistic was not calculated.[16]

Model development was performed using the 2012-2013 data from ACS-NSQIP-P. The resulting SL was then validated on the 2014 ACS-NSQIP-P dataset. Both discrimination and calibration were assessed in the validation dataset. To evaluate whether a smaller number of predictors might yield an SL estimator with similar performance as that produced using the full set of preoperative variables and procedures, all analyses were rerun using only those predictors associated with mortality at a *P*-value of less than 0.20 in bivariate analyses that consisted of Wilcoxon rank sum tests and Pearson chi square tests. To even further reduce model size, and potentially improve model performance on external validation, two additional SL estimators were calculated after 1) eliminating all indicator variables for procedures (CPT codes) and 2) including only those predictors associated with mortality at a *P*-value of less than 0.01 in bivariate analyses. Finally, the performance of the original SL and each SL that considered only subsets of the predictors was assessed specifically in preterm neonates, as these infants are at a higher risk of postoperative death than term neonates. All analyses were performed with SAS version 9.4 (SAS Institute Inc, Cary, NC) or R version 3.3.0 (R Foundation for Statistical Computing, Vienna, Austria). In R, the SuperLearner,[17] cvAUROC,[13] pROC,[18] and rms[19] packages were used.

Results

Population characteristics

A total of 6499 neonatal surgical cases were included in ACS-NSQIP-P in 2012 and 2013. More than 60% of the neonates were male and over half were born preterm. Most preoperative characteristics varied significantly between patients who died within 30 days of surgery and patients who survived (Table 1). The most common procedures performed in the neonates, stratified by whether the patient died within 30 days of the operation, are shown in Table 2. Principal procedures in at least 2% of the patients in each group are shown. Among neonates who died within 30 days of surgery, more than 20% had undergone an exploratory laparotomy.

Model development and performance

Figure 1 shows the performance of the SL and each of its 14 candidate algorithms, as assessed by the cross-validated mean squared error. Performance as measured by cross-validated AUROC is shown in Table 3.

As suggested by theory, the SL performed as well as the best of all 14 candidate algorithms with regard to both CV-MSE and CV-AUROC. The discrimination of the superlearner was excellent, with a CV-AUROC of 0.91. Most of the candidate algorithms also had excellent discrimination, with those that performed well ranging from 0.87 for stepwise regression with forward selection to 0.91 for a random forest with 50 variables randomly selected for consideration at each node and a minimum node size of either 1 or 5. Only single classification trees had poor discrimination, with CV-AUROCs of only 0.59. The SL showed good calibration in the data in which it was developed (Fig. 2). At 0.36 and 1.16, the calibration intercept and slope were close to their optimal values of 0 and 1 respectively (U statistic 0.0006, $P = 0.06$). The calibration of each of the candidate algorithms was also good (data not shown). When a second SL model was fit, with predictors restricted to just those associated with mortality at $P < 0.20$ in bivariate analyses, the performance of this SL model was similar to that of the original SL. The CV-AUROC was 0.91 (95% CI 0.89-0.93), and the Cox calibration test yielded U = -0.0004, $P = 0.09$. A third SL model, fit without any indicator variables for procedures, showed nearly identical performance. The CV-AUROC was 0.91 (95% CI 0.90-0.93), and the Cox calibration test yielded U = -0.0004, $P = 0.09$. When a fourth SL estimator was calculated, with predictors restricted to just those associated with mortality at $P < 0.01$ in bivariate analyses, its performance was also similar to that of the original SL. The CV-AUROC was 0.91 (95% CI 0.89-0.93), and the Cox calibration test yielded U = -0.0003, $P = 0.11$.

Figure 3 shows the observed and predicted neonatal postoperative mortality rates by gestational age at birth for neonates

Table 1 – Preoperative characteristics of neonates treated in 2012-2013.

Characteristic	All patients ($n = 6499$)	Survived ($n = 6267$)	Died ($n = 232$)	P-value
Age at surgery (days)	16 (3-50)	16 (3-53)	14 (7-28)	0.47
Gestational age at birth (weeks)		39 (36-41)	33 (29-38)	<0.001
<24	110 (1.7)	98 (1.6)	12 (5.2)	
24	219 (3.4)	197 (3.1)	22 (9.5)	
25-26	499 (7.7)	457 (7.3)	42 (18.1)	
27-28	333 (5.1)	311 (5.0)	22 (9.5)	
29-30	300 (4.6)	278 (4.4)	22 (9.5)	
31-32	361 (5.6)	348 (5.6)	13 (5.6)	
33-34	563 (8.7)	542 (8.6)	21 (9.1)	
35-36	1024 (15.8)	1004 (16.0)	20 (8.6)	
≥37	3090 (47.5)	3032 (48.4)	58 (25.0)	
Female	2508 (38.6)	2405 (38.4)	103 (44.4)	0.06
Race				0.02
White	4172 (64.2)	4041 (64.5)	131 (56.5)	
Black	1021 (15.7)	968 (15.4)	53 (22.8)	
Asian	141 (2.2)	136 (2.2)	5 (2.2)	
Other/unknown	1165 (17.9)	1122 (17.9)	43 (18.5)	
Birth weight (kg)*	2.48 (1.38-3.14)	2.50 (1.46-3.15)	1.25 (0.78-2.39)	<0.001
Weight at surgery (kg)†	3.07 (2.39-3.71)	3.09 (2.43-3.74)	1.60 (1.06-2.92)	<0.001
Ventilator dependent	1704 (26.2)	1503 (24.0)	201 (86.6)	<0.001
BPD or chronic lung disease	1035 (15.9)	978 (15.6)	57 (24.6)	<0.001
Oxygen support	1825 (28.1)	1656 (26.4)	169 (72.8)	<0.001
Structural pulmonary/airway abnormality	868 (13.4)	814 (13.0)	54 (23.3)	<0.001
Esophageal gastrointestinal disease	4053 (62.4)	3897 (62.2)	156 (67.2)	0.12
Hepatobiliary pancreatic disease	268 (4.1)	247 (3.9)	21 (9.1)	<0.001
Cardiac risk factors				<0.001
None	4037 (62.1)	3928 (62.7)	109 (47.0)	
Minor	1281 (19.7)	1240 (19.8)	41 (17.7)	
Major	1032 (15.9)	969 (15.5)	63 (27.2)	
Severe	149 (2.3)	130 (2.1)	19 (8.2)	
Structural CNS abnormality	1166 (17.9)	1129 (18.0)	37 (15.9)	
Any congenital malformation	2938 (45.2)	2855 (45.6)	83 (35.8)	
Most common congenital malformations				
Gastroschisis	225 (3.5)	222 (3.5)	3 (1.3)	0.07
Atresia of colon, rectum, or anus	216 (3.3)	209 (3.3)	7 (3.0)	0.79
Atresia of small intestine	181 (2.8)	178 (2.8)	3 (1.3)	0.16
Congenital hypertrophic pyloric stenosis	168 (2.6)	168 (2.7)	0 (0.0)	0.01
Down's syndrome	168 (2.6)	160 (2.6)	6 (2.6)	0.97
Open wound	750 (11.5)	732 (11.7)	18 (7.8)	0.07
Steroid use in previous 30 days	395 (6.1)	343 (5.5)	52 (22.4)	<0.001
Nutritional support	2851 (43.9)	2691 (42.9)	160 (69.0)	<0.001
Hematologic disorder	862 (13.3)	784 (12.5)	78 (33.6)	<0.001
Inotropic support	382 (5.9)	284 (4.5)	98 (42.2)	<0.001
CPR or ECMO in previous 7 days	103 (1.6)	79 (1.3)	24 (10.3)	<0.001
Transfusion in previous 48 hours	568 (8.7)	478 (7.6)	90 (38.8)	<0.001
Intraventricular hemorrhage				0.001
None	5720 (88.0)	5536 (88.3)	184 (79.3)	
Grade 1	205 (3.2)	190 (3.0)	15 (6.5)	
Grade 2	125 (1.9)	116 (1.9)	9 (3.9)	

(continued)

Table 1 – (*continued*)

Characteristic	All patients ($n = 6499$)	Survived ($n = 6267$)	Died ($n = 232$)	*P*-value
Grade 3	143 (2.2)	135 (2.2)	8 (3.4)	
Grade 4	247 (3.8)	235 (3.7)	12 (5.2)	
Unknown grade	59 (0.9)	55 (0.9)	4 (1.7)	
Sepsis in previous 48 hours				<0.001
No	6179 (95.1)	6013 (95.9)	166 (71.6)	
SIRS	73 (1.1)	65 (1.0)	8 (3.4)	
Sepsis	149 (2.3)	125 (2.0)	24 (10.3)	
Septic shock	98 (1.5)	64 (1.0)	34 (14.7)	
ASA class				<0.001
ASA-1	360 (5.5)	360 (5.7)	0 (0.0)	
ASA-2	1482 (22.8)	1479 (23.6)	3 (1.3)	
ASA-3	3082 (47.4)	3049 (48.7)	33 (14.2)	
ASA-4	1423 (21.9)	1286 (20.5)	137 (59.1)	
ASA-5	114 (1.8)	60 (1.0)	54 (23.3)	
Unknown	38 (0.6)	33 (0.5)	5 (2.2)	

BPD = bronchopulmonary dysplasia; CPR = cardiopulmonary resuscitation; ECMO = extracorporeal membrane oxygenation; ASA = American Society of Anesthesiologists.
* $n = 5515$.
† $n = 6464$.

treated in 2012-2013. Neonates who were born more prematurely had a higher mortality risk in general, but after accounting for all available preoperative characteristics, there was a wide range of predicted probabilities of mortality within each gestational age category. The mean predicted probability of mortality was very close to the observed mortality rate in each group.

Model validation

A total of 3552 neonatal surgical cases were included in ACS-NSQIP-P in 2014. As was the case in the patients treated in 2012-2013, approximately 60% of the neonates were male and just over half were born preterm. All other preoperative characteristics of the 2014 patients were also similar to those of the 2012-2013 patients. Most preoperative characteristics varied significantly between patients who did and did not survive 30 days postsurgery (data not shown). Although the procedure mix varied little in 2012-2014 among patients who had more common procedures, there were 65 distinct procedures (CPT codes) in 2012-13 patients that were not present in 2014 patients. These procedures were associated with a similar overall mortality rate as the procedures that were present in the 2014 data

Table 2 – Most common procedures by 30-day mortality status.

Principal procedure in patients that survived ($n = 6267$)	N (%)	Principal procedure in patients that died ($n = 232$)	N (%)
Pyloromyotomy	687 (11.0)	Exploratory laparotomy	47 (20.3)
Creation of VP shunt	362 (5.8)	Enterectomy with enterostomy	32 (13.8)
Repair of large omphalocele or gastroschisis	329 (5.3)	Congenital diaphragmatic hernia repair	18 (7.8)
Laparoscopic gastrostomy	319 (5.1)	Repair of large omphalocele or gastroschisis	11 (4.7)
Ladd procedure	215 (3.4)	Enterectomy, single resection, and anastomosis	10 (4.3)
Congenital diaphragmatic hernia repair	213 (3.4)	Partial colectomy, with colostomy or ileostomy and creation of mucofistula	9 (3.9)
Laparoscopic Nissen fundoplication	203 (3.2)	Colostomy	7 (3.0)
Enterectomy with enterostomy	203 (3.2)	Planned tracheostomy	6 (2.6)
Esophagoplasty, thoracic approach, with repair of congenital tracheoesophageal fistula	183 (2.9)	Drainage of peritoneal abscess or localized peritonitis	6 (2.6)
Planned tracheostomy	183 (2.9)		
Closure of enterostomy with resection and anastomosis other than colorectal	178 (2.8)		
Enterectomy, single resection, and anastomosis	158 (2.5)		
Colostomy	143 (2.3)		
Open gastrostomy	138 (2.2)		

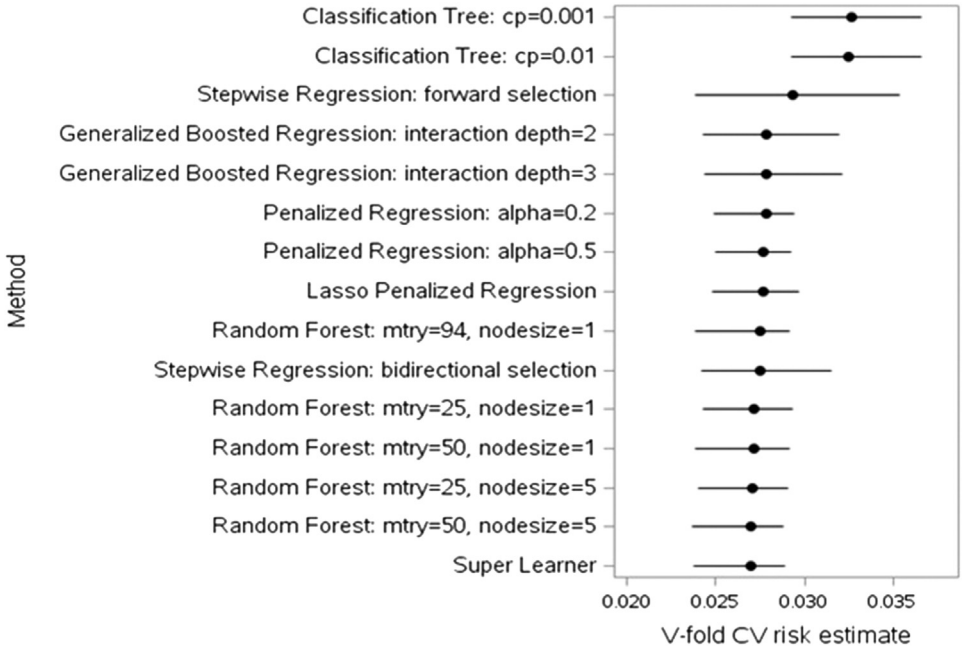

Fig. 1 – Cross-validated mean squared error for superlearner and 14 candidate algorithms. cp = complexity parameter; alpha = elastic-net mixing parameter.

Table 3 – Cross-validated AUROC for superlearner and 14 candidate algorithms.

Algorithm	Cross-validated AUROC (95% CI)
Classification tree: $c_p = 0.001$	0.59 (0.51-0.66)
Classification tree: $c_p = 0.01$	0.59 (0.51-0.66)
Stepwise regression: forward selection	0.87 (0.84-0.90)
Generalized boosted regression: interaction depth = 2	0.90 (0.89-0.93)
Generalized boosted regression: interaction depth = 3	0.90 (0.89-0.93)
Penalized regression: $\alpha = 0.2$	0.90 (0.88-0.92)
Penalized regression: $\alpha = 0.5$	0.90 (0.88-0.92)
Lasso-penalized regression	0.90 (0.88-0.92)
Random forest: mtry = 94, node size = 1	0.90 (0.88-0.93)
Stepwise regression: bidirectional selection	0.89 (0.87-0.92)
Random forest: mtry = 25, node size = 1	0.89 (0.87-0.92)
Random forest: mtry = 50, node size = 1	0.91 (0.89-0.93)
Random forest: mtry = 25, node size = 5	0.89 (0.87-0.92)
Random forest: mtry = 50, node size = 5	0.91 (0.89-0.93)
Superlearner	0.91 (0.89-0.93)

c_p = complexity parameter; α = elastic-net mixing parameter; mtry = number of variables randomly selected for consideration at each split.

set (4.0% vs. 3.6%), and they accounted for only 1.9% of the procedures performed in 2012-13. There were also 59 distinct procedures in the 2014 data set that were not found in the 2012-13 data set; these accounted for 4.5% of the procedures performed in 2014 but were associated with only 1 death.

The discrimination of the SL remained excellent in the 2014 data set (AUROC = 0.87, 95% CI 0.83-0.90). Its calibration in 2014 patients, however, was poor (U statistic 0.064, $P < 0.001$). The SL model overestimated mortality risk among patients at a low risk of mortality but underestimated mortality risk among patients at the highest risk (Fig. 4). The calibration of each candidate algorithm in the SL was also poor, with the

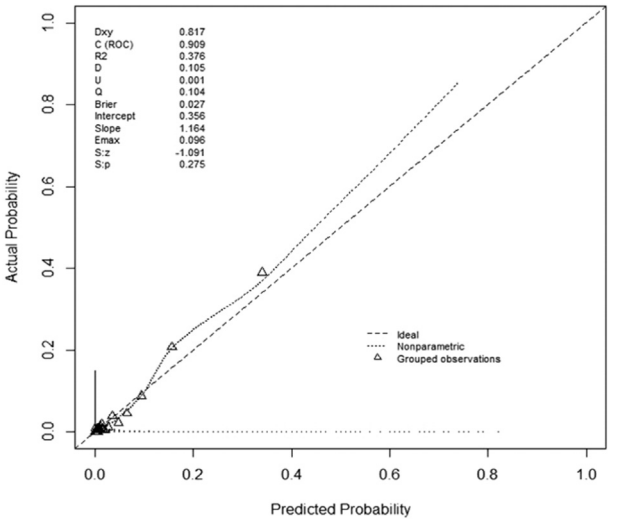

Fig. 2 – Calibration plot for superlearner in development data set.

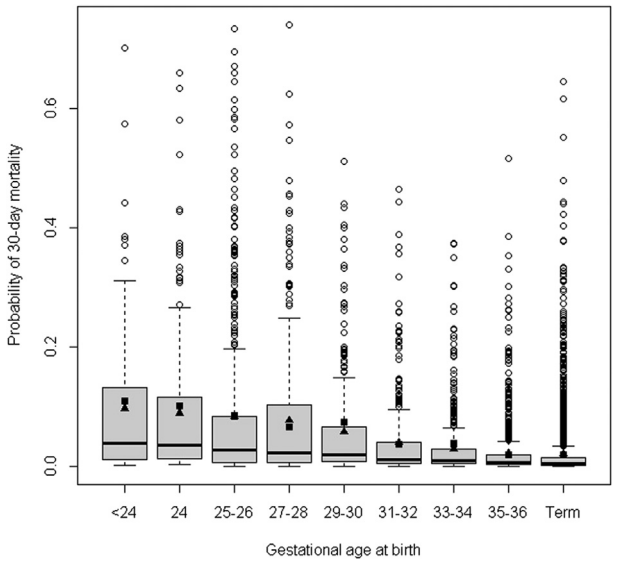

Fig. 3 – Boxplot of predicted probabilities of mortality by gestational age. ▲ = mean predicted probability of mortality; ■ = observed mortality rate.

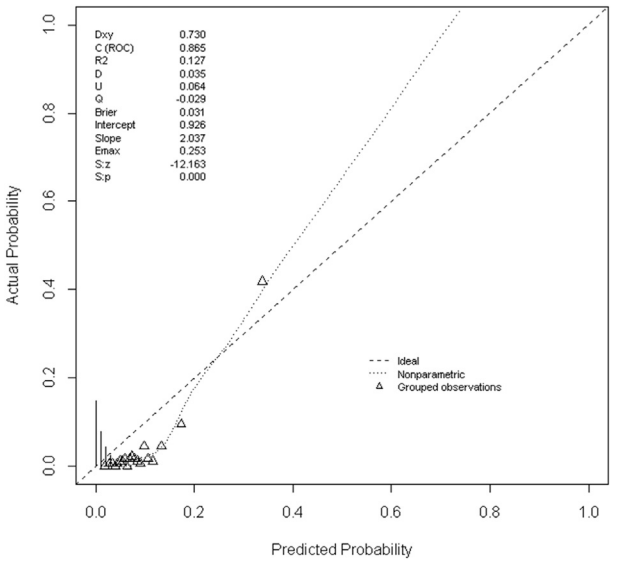

Fig. 4 – Calibration plot for superlearner in validation data set.

exception of the boosted CART algorithms (Cox calibration test $P > 0.20$ for both).

In preterm neonates specifically, the discrimination of the SL was again excellent (AUROC = 0.85, 95% CI 0.81-0.89). Its calibration was slightly improved from that in the total sample but still poor (U statistic = 0.049, $P < 0.001$). When just the patients with procedures that were performed in at least 10 cases in both 2012-13 and 2014 were included in the validation dataset ($n = 3014$), the SL model's discrimination remained excellent (AUROC = 0.86, 95% CI 0.82-0.90) but its calibration remained poor (U statistic = 0.064, $P < 0.001$). Results were similar when only patients with procedures performed in at least 1% of cases during both time periods were included in validation (AUROC = 0.86, 95% CI 0.82-0.91; U statistic = 0.063, $P < 0.001$). When the SL model that included only predictors associated with mortality at $P < 0.20$ in bivariate analysis was tested, its performance was similar to that of the original SL in the 2014 data set. The AUROC was 0.87 (95% CI 0.83-0.91), and the Cox calibration test yielded U = 0.068, $P < 0.001$. When the SL model that did not include indicator variables for procedures was tested, its performance was similar to that of the original SL in the 2014 data set. The AUROC was 0.86 (95% CI 0.82-0.89), and the Cox calibration test yielded U = 0.072, $P < 0.001$. Finally, when the SL estimator including only the 57 predictors that were associated with mortality at $P < 0.01$ in bivariate analysis was tested, its performance was superior to that of the original SL in the validation data set. Though it showed similar discrimination (AUROC 0.89, 95% CI 0.86-0.92), its calibration was excellent (U = −0.0003, $P = 0.63$). Interestingly, all component algorithms of this SL also showed good calibration, with the exception of the logistic regression model fit using forward variable selection (U = 0.016, $P < 0.001$). Performance was similar when only preterm neonates were evaluated (data not shown).

Discussion

This study showed that superlearning provided improved or equivalent accuracy compared with individual regression and machine learning algorithms for predicting neonatal surgical mortality. Because the optimal prediction algorithm in any data set cannot be known unless the data generating mechanism is also known, data-adaptive algorithms such as superlearning can be helpful for choosing among a variety of candidate prediction algorithms when the data generating mechanism is poorly understood. Superlearning offers a flexible alternative to other nonparametric methods because it can include as many candidate algorithms as desired and will perform at least as well as the best individual candidate algorithm in its library. It should be considered for prediction in large data sets whenever complex mechanisms make parametric modeling assumptions unrealistic.

In this study, our SL model had a similar or higher AUROC than previously published studies that used logistic regression to predict surgical mortality or morbidity in neonates undergoing a variety of noncardiac surgical procedures.[3,4,20] However, although an SL will perform no worse than its best constituent algorithm in a training data set, as with any model there is no guarantee that it will perform well in a validation data set. Poor calibration of the original SL in our validation data set may have been partly due to the addition of 11 new hospitals, and many new surgeons at those hospitals, to ACS-NSQIP-P in 2014. Numerous studies in both adult and pediatric surgery have shown that hospital and surgeon characteristics are associated with patients' outcomes after surgery.[21,22] For example, hospital experience with a procedure strongly predicts outcomes after surgery for biliary atresia and congenital diaphragmatic hernia.[23,24] The level of care and experience with very low birth weight neonates of the neonatal intensive care unit at the delivery hospital also influence a neonate's risk of mortality and morbidity.[25,26] Unfortunately there are

no surgeon or hospital-level identifiers or characteristics in the ACS-NSQIP-P PUF; therefore we were not able to account for these factors. However, given the excellent calibration of the SL model that included only 57 preoperative variables strongly associated with mortality in bivariate analyses, it appears that the poor calibration of the original SL model was primarily due to overfitting. As there were only 232 deaths within 30 days of surgery among neonates in 2012-2013, our consideration of 284 predictors likely resulted in substantial overfitting. In high-dimensional data, it is often necessary to screen variables before fitting any type of multivariable model, whether a regression model or an ensemble of machine learning algorithms such as the superlearner.

Unlike some previous studies that have reported on the performance of logistic regression models for predicting mortality after pediatric surgical procedures,[1,2] this study performed a thorough evaluation of model calibration. Although AUROCs are nearly universally provided for prediction models reported in the literature, a high AUROC, though indicative of good model discrimination, does not necessarily imply good overall model fit,[27] particularly when the outcome is rare as was the case in this study. Therefore, it is important to evaluate other measures of model performance during model development and validation.

Although the SL may seem to be a "black box" algorithm, from which the effects of individual variables on the outcome are not easily discerned, such as with the odds ratios available from logistic regression, measures of variable importance can be calculated after model fitting. Targeted maximum likelihood estimators, which provide an estimate of the average effect a particular exposure would have on the outcome in the entire sample, can be calculated after an SL model is fit.[28] However, such marginal effects, just like the conditional effects produced by a logistic regression model, have little practical meaning in the present study because the effects of many of the available modifiable patient characteristics, such as prematurity for example, are mediated through other characteristics that are also predictors in the SL, such as birth weight, preoperative mechanical ventilation, bronchopulmonary dysplasia, and nutritional support. Furthermore, when the primary goal of an analysis is to achieve optimal prediction rather than effect estimation, the lack of "clinical interpretability" of an algorithm should not be considered a disadvantage. It must be noted, however, that the superlearner algorithm is more computationally intensive than standard regression modeling. It is recommended that the superlearner be composed of a large collection of component algorithms that vary in their bias and degree of data-fitting.[28] In fact, its performance can only be improved by adding more component algorithms. However, fitting such a model may take hours or even days to run on a typical personal computer. Fortunately, the superlearner algorithm is perfectly tailored for parallel programming as the component algorithms can be run separately and the applications of the algorithms to each training set in cross-validation can also be separated, greatly reducing computation time.

This study had several limitations. First, there was a substantial amount of missing data for some variables, particularly preoperative lab values. Although we included all available neonatal cases and did not exclude those with missing data, we did not perform multiple imputations to fill in missing data because we doubted that the missing lab values could be reliably imputed using the available clinical patient characteristics. This could have introduced some bias into the SL model. Second, no condition-specific patient characteristics are available in the ACS-NSQIP-P PUF. Although this is necessary for the development of widely applicable models for risk adjustment and outcome prediction, it does potentially limit the accuracy of such models for particular patient populations. Another limitation of this study was our accounting for procedures by treating individual principal and concurrent procedures as distinct binary variables. This prevented us from incorporating all procedures as distinct predictors because of the rarity of some procedures. This method also differs from that used by the American College of Surgeons (ACS) to account for procedure mix, which relies on the use of a CPT linear risk score.[29] The groupings of procedures used by the ACS to calculate this risk score, however, are not publicly available. Other investigators have used work relative value units, which are indicative of the time and effort required of the surgeon to perform the procedure. In addition, our inclusion of individual procedures as indicator variables, as well as our inclusion of a large number of other variables, led to event per variable ratios of only 1-2. This is far less than the 10-20 events per variable that is generally recommended for the fitting of prediction models.[30,31] Thus, it is likely that our algorithms were somewhat overfit to the data in which they were developed, which contributed to their poor calibration in the validation data set. The improved calibration of the algorithm that was fit using only predictors highly significantly associated with mortality in bivariate analysis strongly suggests the overfitting of the original SL. Finally, as previously mentioned, the list of procedures included in ACS-NSQIP-P changes slightly from year to year, which contributed to the differences in procedure mix between our development and validation data sets. This temporal change in procedure mix appears to have had minimal effect on our model validation as demonstrated by the minimal changes in discrimination and calibration of the SL when the validation data set was restricted to patients whose procedures were common in both the 2012-13 and 2014 data sets. Regarding procedure mix, it should be noted that ACS-NSQIP-P does not include some common low risk surgical procedures, such as circumcision, myringotomy, and inguinal hernia repair. Thus, the mortality rate reported in this study is likely an overestimate of the true mortality rate among all neonatal noncardiac surgery patients. In addition, the SL model developed in this study would likely inaccurately predict mortality in patients with procedures not included in ACS-NSQIP-P.

Conclusions

In conclusion, superlearning provided similar or superior accuracy and discrimination compared with individual regression and machine learning algorithms for predicting neonatal surgical mortality. Given the complex and variable processes underlying operative mortality in neonates, this flexible statistical method may yield more reliable mortality predictions in these patients than standard parametric regression models.

Acknowledgment

The American College of Surgeons National Surgical Quality Improvement Program and the hospitals participating in the ACS NSQIP Pediatric are the source of the data used herein; they have not verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors.

Author's contributions: Dr. Cooper designed the study, performed all statistical analysis, interpreted the data, and wrote the initial draft of the manuscript. Dr. Minneci interpreted the data and revised the manuscript. Dr. Deans interpreted the data and revised the manuscript. All authors approve of the final version of the manuscript as submitted.

Disclosure

The authors report no proprietary or commercial interest in any product mentioned or concept discussed in this article.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jss.2017.09.002.

REFERENCES

1. Bucher BT, Duggan EM, Grubb PH, France DJ, Lally KP, Blakely ML. Does the American College of Surgeons National Surgical Quality Improvement Program pediatric provide actionable quality improvement data for surgical neonates? *J Pediatr Surg*. 2016;51:1440–1444.
2. Langham Jr MR, Walter A, Boswell TC, Beck R, Jones TL. Identifying children at risk of death within 30 days of surgery at an NSQIP pediatric hospital. *Surgery*. 2015;158:1481–1491.
3. Stey AM, Kenney BD, Moss RL, et al. A risk calculator predicting postoperative adverse events in neonates undergoing major abdominal or thoracic surgery. *J Pediatr Surg*. 2015;50:987–991.
4. Lillehei CW, Gauvreau K, Jenkins KJ. Risk adjustment for neonatal surgery: a method for comparison of in-hospital mortality. *Pediatrics*. 2012;130:e568–e574.
5. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6. Article25.
6. Bruny JL, Hall BL, Barnhart DC, et al. American College of Surgeons National Surgical Quality Improvement Program pediatric: a beta phase report. *J Pediatr Surg*. 2013;48:74–80.
7. Dillon P, Hammermeister K, Morrato E, et al. Developing a NSQIP module to measure outcomes in children's surgical care: opportunity and challenge. *Semin Pediatr Surg*. 2008;17:131–140.
8. Raval MV, Dillon PW, Bruny JL, et al. American College of Surgeons National Surgical Quality Improvement Program pediatric: a phase 1 report. *J Am Coll Surg*. 2011;212:1–11.
9. Breiman L. *Classification and regression trees*. Belmont, Calif: Wadsworth International Group; 1984.
10. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22.
11. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29:1189–1232.
12. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38:367–378.
13. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron J Stat*. 2015;9:1583–1607.
14. Cox DR. Two further applications of a model for binary regression. *Biometrika*. 1958;45:562–565.
15. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. New York, NY: Springer; 2009.
16. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. *Crit Care Med*. 2007;35:2052–2056.
17. Polley EC, van der Laan MJ, SuperLearner: Super Learner Prediction, R package version 2.0-15. [computer program], 2014. Available at: https://cran.r-project.org/web/packages/SuperLearner/.
18. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
19. Harrell FE, rms: Regression Modeling Strategies, R package version 4.1-0. [computer program], 2013. Available at: https://cran.r-project.org/web/packages/rms/.
20. Son JK, Lillehei CW, Gauvreau K, Jenkins KJ. A risk adjustment method for newborns undergoing noncardiac surgery. *Ann Surg*. 2010;251:754–758.
21. Reames BN, Ghaferi AA, Birkmeyer JD, Dimick JB. Hospital volume and operative mortality in the modern era. *Ann Surg*. 2014;260:244–251.
22. McAteer JP, LaRiviere CA, Drugas GT, Abdullah F, Oldham KT, Goldin AB. Influence of surgeon experience, hospital volume, and specialty designation on outcomes in pediatric surgery: a systematic Review. *JAMA Pediatr*. 2013;167:468–475.
23. McKiernan PJ, Baker AJ, Kelly DA. The frequency and outcome of biliary atresia in the UK and Ireland. *Lancet*. 2000;355:25–29.
24. Grushka JR, Laberge JM, Puligandla P, Skarsgard ED. Canadian Pediatric Surgery N. Effect of hospital case volume on outcome in congenital diaphragmatic hernia: the experience of the Canadian Pediatric Surgery Network. *J Pediatr Surg*. 2009;44:873–876.
25. Phibbs CS, Baker LC, Caughey AB, Danielsen B, Schmitt SK, Phibbs RH. Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. *N Engl J Med*. 2007;356:2165–2175.
26. Jensen EA, Lorch SA. Effects of a birth Hospital's neonatal intensive care unit level and annual volume of very low-birth-weight infant deliveries on morbidity and mortality. *JAMA Pediatr*. 2015;169:e151906.
27. Lobo JM, Jimenez-Valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr*. 2008;17:145–151.
28. van der Laan MJ, Rose S. *Targeted learning: causal inference for observational and experimental data*. New York: Springer; 2011.
29. Raval MV, Cohen ME, Ingraham AM, et al. Improving American College of Surgeons National Surgical Quality Improvement Program risk adjustment: incorporation of a novel procedure risk score. *J Am Coll Surg*. 2010;211:715–723.
30. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373–1379.
31. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol*. 2016;76:175–182.