


Research and Applications

Predicting pediatric emergence delirium using data-driven machine learning applied to electronic health record dataset at a quaternary care pediatric hospital

Han Yu, PhD^{1,2}, Allan F. Simpao, MD, MBI^{3,4}, Victor M. Ruiz, PhD⁴, Olivia Nelson, MD³, Wallis T. Muhly, MD³, Tori N. Sutherland, MD, MPH³, Julia A. Gálvez, MD, MBI⁵, Mykhailo B. Pushkar, MD, PhD⁶, Paul A. Stricker, MD³, Fuchiang (Rich) Tsui , PhD^{*,3,4}

¹Department of Anesthesiology and Critical Care Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, United States, ²Department of Population Medicine, Harvard Medical School & Harvard Pilgrim Health Care Institute, Boston, MA 02215, United States, ³Department of Anesthesiology and Critical Care Medicine, The Children's Hospital of Philadelphia and the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, United States, ⁴Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, United States, ⁵Department of Anesthesiology & Critical Care, Children's Hospital & Medical Center, Omaha, NE 68114, United States, ⁶Department of Anesthesiology, Intensive Care and Pediatric Anesthesiology, Kharkiv National Medical University, Kharkiv, 61022, Ukraine

*Corresponding author: Fuchiang (Rich) Tsui, PhD, Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, 2716 South St, Philadelphia, PA 19146 (tsuif@chop.edu)

Abstract

Objectives: Pediatric emergence delirium is an undesirable outcome that is understudied. Development of a predictive model is an initial step toward reducing its occurrence. This study aimed to apply machine learning (ML) methods to a large clinical dataset to develop a predictive model for pediatric emergence delirium.

Materials and Methods: We performed a single-center retrospective cohort study using electronic health record data from February 2015 to December 2019. We built and evaluated 4 commonly used ML models for predicting emergence delirium: least absolute shrinkage and selection operator, ridge regression, random forest, and extreme gradient boosting. The primary outcome was the occurrence of emergence delirium, defined as a Watcha score of 3 or 4 recorded at any time during recovery.

Results: The dataset included 54 776 encounters across 43 830 patients. The 4 ML models performed similarly with performance assessed by the area under the receiver operating characteristic curves ranging from 0.74 to 0.75. Notable variables associated with increased risk included adenoidectomy with or without tonsillectomy, decreasing age, midazolam premedication, and ondansetron administration, while intravenous induction and ketorolac were associated with reduced risk of emergence delirium.

Conclusions: Four different ML models demonstrated similar performance in predicting postoperative emergence delirium using a large pediatric dataset. The prediction performance of the models draws attention to our incomplete understanding of this phenomenon based on the studied variables. The results from our modeling could serve as a first step in designing a predictive clinical decision support system, but further optimization and validation are needed.

Clinical trial number and registry URL: Not applicable.

Lay Summary

Pediatric emergence delirium is a transient phenomenon that occurs in children as they wake up (emerge) from anesthesia in which they may have disturbances in awareness of and attention to their environment, disorientation, hypersensitivity to stimuli, and hyperactive motor behaviors. Emergence delirium is an undesirable outcome whose accurate prediction could allow clinicians to administer targeted preventive therapy. This study applied machine learning methods to a large clinical dataset to develop a predictive model for pediatric emergence delirium. The dataset included 54 776 encounters across 43 830 patients. The models tested had moderate predictive performance, drawing attention to our incomplete understanding of this phenomenon. Several variables were identified to be associated with an increased risk of emergence delirium, while others were identified to be associated with a reduced risk of emergence delirium. The results from our modeling could serve as a first step in designing a predictive clinical decision support system, but further optimization and validation are needed.

Implication statement

Four machine learning predictive models for pediatric emergence delirium were developed and tested using a large dataset, with nearly indistinguishable performance (area under the receiver operating characteristic curve: 0.74-0.75). Prospective study is required to assess whether model-guided interventions meaningfully reduce delirium.

Introduction

Pediatric emergence delirium is a transient postoperative phenomenon in which children may have disturbances in awareness of and attention to their environment, disorientation, hypersensitivity to stimuli, and hyperactive motor behaviors.¹ The reported incidence ranges widely¹⁻⁸ and varies depending on the measurement instrument used and the population assessed. Although emergence delirium usually resolves within thirty minutes,^{1,9} it may cause significant distress. Children may injure themselves or staff, and they may dislodge catheters and surgical drains.¹⁰⁻¹² Management often requires additional personnel and/or treatment, and prolongation of the postoperative care unit (PACU) stay is common. Furthermore, there is emerging data that emergence delirium is associated with negative behavioral changes beyond the immediate postoperative period.^{13,14}

Development of a method of identifying children at risk for emergence delirium could allow for the development of clinical decision support to help reduce the incidence of this outcome. Nearly all the known factors associated with pediatric emergence delirium were identified through studies with relatively small sample sizes, with few describing predictive model development.¹⁵⁻¹⁷ The generalizability of controlled trials and observational studies can be unclear in daily clinical practice. Many of these studies are based on carefully selected populations, whereas in practice clinicians confront heterogeneous populations in which the applicability of such data is uncertain. When applied to large training datasets, supervised data-driven machine learning (ML) algorithms using electronic health record (EHR) data may provide accurate outcome prediction with minimal human guidance.¹⁸ Although more recent studies have started using ML algorithms to predict emergence delirium, they mainly focus on adult populations, use none or limited intra-operative variables (eg, anesthesia maintenance and medications), and measure predictive performance at one time-point such as pre-operation.¹⁹⁻²¹

The routine inclusion of management and outcome data in the EHR has produced large observational datasets for analysis. The primary aim of this study was to apply ML methods to a large quaternary institution's EHR dataset to develop a predictive model for pediatric emergence delirium. The secondary aim was to determine the adjusted and unadjusted odds ratios for emergence delirium of the study variables using multivariate and univariate analysis, respectively. We hypothesize that data-driven ML models can accurately predict pediatric emergence delirium using a large retrospective pediatric cohort. Our contribution includes (1) large comprehensive pediatric EHR data collected at 3 perioperative time points: pre-operation, intra-operation, and post-operation, (2) development of 4 ML models, (3) predictive model evaluation at 3 perioperative time points, and (4) individual study variables analysis using adjusted and unadjusted odds ratios.

Materials and methods

Ethics approval and reporting guidelines

The Children's Hospital of Philadelphia Institutional Review Board (IRB) reviewed this study and granted exemption status and a waiver for written informed consent on November 21, 2019 (IRB 19-016984); the study abides by the Ethical Principles for Medical Research Involving Human Subjects outlined in the Declaration of Helsinki. The manuscript adheres to the Transparent Reporting of a multivariable Prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.²² The study analysis plan was developed prior to accessing the data.

Patient population

We performed a retrospective cohort study using prospectively entered data in the EHR (Epic Systems Corporation, Verona, WI, United States) from patients receiving anesthetic care from February 1, 2015 to December 31, 2019 at the Children's Hospital of Philadelphia. The study inclusion and exclusion criteria, variable definitions, and analysis plan were established before the analysis. Study inclusion criteria were age 2 years to <13 years undergoing procedures at the main hospital or ambulatory surgery center. Exclusion criteria were patients having non-operating room anesthesia (eg, interventional radiology, diagnostic radiology), patients who underwent anesthesia in the cardiac suite, patients not recovering in the PACU, and patients missing primary outcome data.

Primary outcome

Assessment and documentation of emergence delirium in our PACU were standardized in 2015 using the Watcha scale,²³ in which children are assessed approximately every 15 min during Phase 1 of recovery as follows: 1 = calm, quiet; 2 = crying, but consolable; 3 = crying, inconsolable; 4 = agitated and thrashing around. The primary outcome was the occurrence of emergence delirium defined as a Watcha score of 3 or 4 recorded at any time point during recovery.²⁴

Study dataset

The EHR database was queried, and data filtering and validation were performed to identify eligible patients. The pediatric emergence delirium literature was reviewed to identify potential variables for inclusion. All variables identified that were available in the EHR were included. The study investigators conferred and selected additional variables for inclusion. Data extracted for each subject included age, weight, height, sex, American Society of Anesthesiologists (ASA) physical status, race, ethnicity, gestational age at birth, history of prior surgery, surgical procedure, International Classification of Diseases codes (to determine history of developmental delay, seizure disorder, asthma/reactive airway disease, obstructive sleep apnea, attention deficit/hyperactivity disorder, and autism spectrum disorder), postoperative disposition (eg, day surgery, admit after surgery), surgical facility type (ambulatory surgery center vs hospital), fasting times, preoperative heart rate, preoperative blood pressure, sedative premedications, patient behavior at induction, parental presence at induction, anesthesia induction type, intraoperative medications, airway management, performance of a regional anesthesia technique, nursing

assessments, and event timestamps. The study variable details are presented in [Supplementary Digital Content Table S1](#). Procedure groupings and categorizations used in the study are presented in [Supplementary Digital Content Tables S2](#) and [S3](#). The schema for categorizing maintenance of anesthesia is described in [Supplementary Digital Content Figures S1](#) and [S2](#).

All eligible patient encounters were included in the study. We conducted the study at the encounter level because there is evidence that the occurrence of emergence delirium can vary in the same subject undergoing the same procedure depending on management,²⁵ and because in our cohort patients underwent different procedures at different ages, both of which influence the likelihood of emergence delirium occurring.³

Missing data and data imputation

For the ML models, missing values for categorical variables were assigned a null value and an indicator variable added (assuming missing not at random), while missing values in continuous variables were populated with the median for that variable in the training dataset. For the multivariable analysis, missing data were managed using multiple imputations. The variables with the highest rates of missingness were behavior at separation (46%), behavior at induction (37%), and induction technique (32%). These rates of missingness occurred because the associated documentation fields were implemented in the EHR after the study start date. Missing values in the dataset were imputed with multiple imputations using the R Hmisc package.²⁶

Machine learning models for predicting emergence delirium

We built and evaluated 4 ML models: least absolute shrinkage and selection operator (LASSO) logistic regression, ridge logistic regression, random forest (RF), and extreme gradient boosting (XGB). We developed the models using a nested, stratified 10-fold cross validation. Nested cross validation is a more rigorous protocol that overcomes the overfitting pitfall of non-nested cross validation.^{27,28} First, we randomly split the dataset into 10 (outer) folds. One of the 10 folds was reserved as a test (hold-out) dataset, and the remaining 9 folds were combined and used as a training dataset. Within the training dataset, a best model was built based on a grid search of best model parameters via (inner) cross validation. We repeated the same process (9-fold training and 1-fold testing) 10 times, and each time a different test dataset (one of the 10 folds) was employed. Model performance was evaluated by the average results from the outer 10-fold cross validation. Model evaluation metrics included the area under the receiver operating characteristics curve (AUROC), area under the precision-recall curve (AUPRC),²⁹ sensitivity, specificity, positive predictive value (PPV), and negative predictive value. The adjusted *P*-values for paired AUROC comparisons were conducted using 2-sided DeLong tests.³⁰

In a secondary analysis using the same methodology, we evaluated the performance of the ML models using the variables that were available at 2 perioperative time points: anesthesia induction and end of surgery. The first time point included data prior to anesthesia induction that would largely not be readily modifiable (eg, demographic data, procedure). The second time point included data through the end of

surgery, and included potentially modifiable features (eg, intraoperative medications).

Measuring unadjusted and adjusted odds ratios of individual variables using regression

To evaluate the impact of individual variables on emergence delirium, we measured unadjusted and adjusted odds ratios of the variables using univariate and multivariable regression analysis from the full dataset without a split of training and testing datasets, respectively. Variables with a *P*-value <.05 in the univariate analysis (unadjusted odds ratios) and those with published evidence supporting an effect on emergence delirium were included in the multivariable model. Bonferroni correction was used to account for multiple comparisons ($\alpha = .05/55$; $P < .0009$ denotes significance). The largest group for each categorical variable was used as the reference in both ML and logistic regression models, except the “no medication documented” group for premedication. The “fit.mult.impute” function in the R package was used to fit all imputations and average the results for the multivariable logistic regression model.³¹ Analysis was performed using Stata/IC 14.2 (College Station, TX: StataCorp, LP), R version 4.0.4 (R Foundation for Statistical Computing, Vienna, Austria) with packages pROC, lrm, Hmisc, and rms, and Python 3.9.5.

Secondary analysis

We performed a secondary analysis by limiting each model to 10 variables to form parsimonious models that could potentially simplify the deployment process using a small number of variables. The evaluation metrics included AUROC, AUPRC, and the Brier skill score. We also performed statistical significance tests for each parsimonious model when compared to its counterpart full-size model.

Results

54 776 encounters in 43 830 patients met the inclusion criteria. A patient cohort flow diagram is presented in [Figure 1](#). Patient characteristics are presented in [Table 1](#). There were 37 135 patients (85%) with a unique encounter, 4740 patients (11%) with 2 encounters, 1085 patients (2.5%) with 3 encounters, and 870 patients (2.0%) with 4 or more encounters. Emergence delirium occurred in 4356 encounters yielding an overall incidence of 8%. A total of 75 variables were retrieved from the EHR data.

The performance of the 4 ML models on the training and test datasets are presented in [Table 2](#). The 4 models exhibited nearly identical performance, as shown in the receiver operating characteristic and precision recall plots in [Figure 2](#). The XGB model outperformed the LASSO model with a slight difference in AUROC of 0.01. However, LASSO was selected for presentation because the minor improvement in XGB is offset by its findings being less interpretable by clinicians as compared to LASSO. The relative feature importance in the LASSO model for the variables with the ten largest magnitude coefficients for increased and decreased likelihood of emergence delirium is presented in [Figure 3](#). The coefficients for all variables in the LASSO model are available in [Supplementary Digital Content Table S4](#). The model predictions for 4 clinical scenarios are presented in [Figure 4](#). In the secondary analysis, the AUROCs of the LASSO regression model at anesthesia induction, end of surgery, and PACU time points

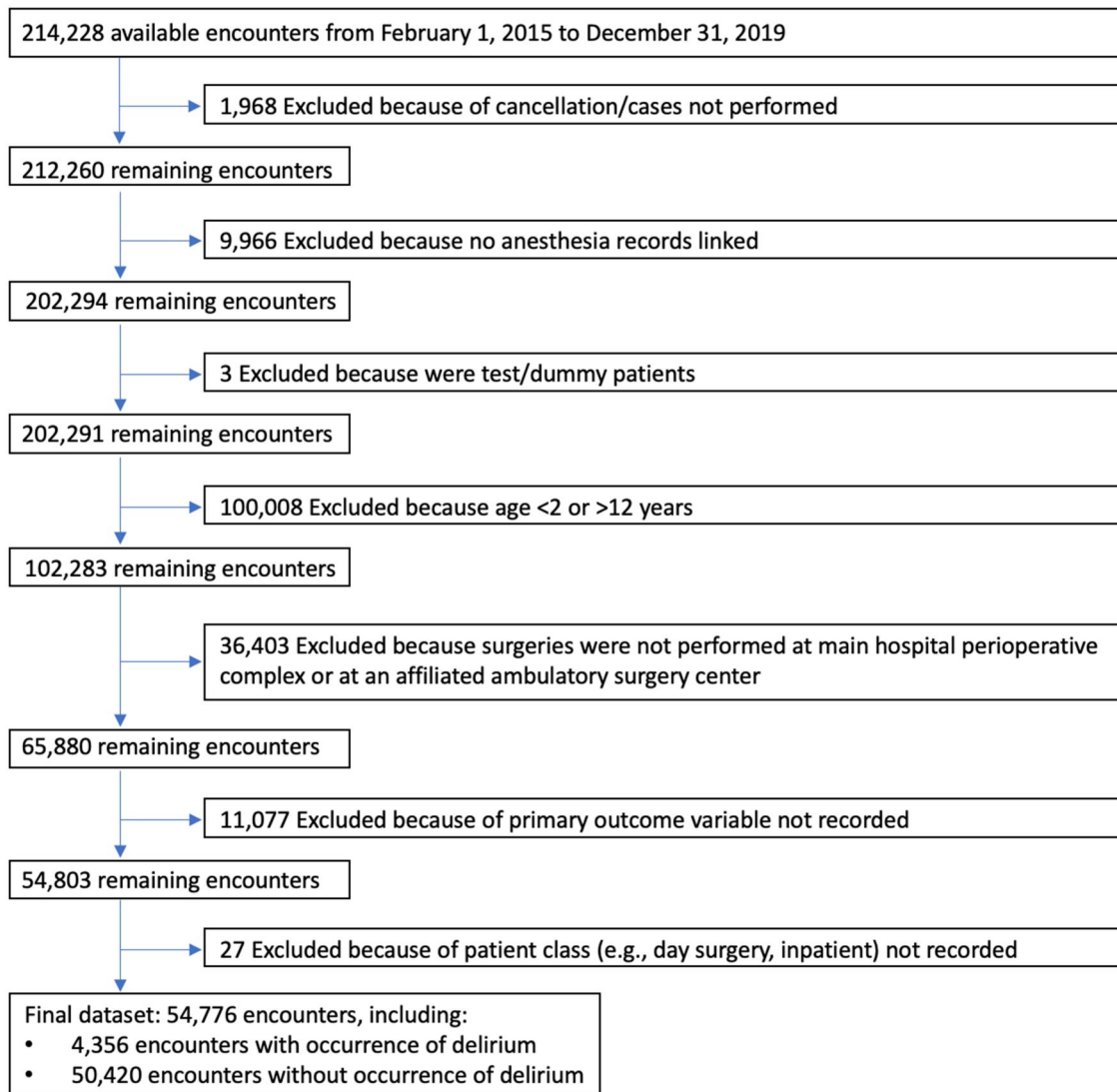


Figure 1. Study flow diagram.

were 0.72 (95% confidence interval [CI]: 0.70-0.75), 0.73 (95%CI: 0.70-0.75), and 0.74 (95%CI: 0.72-0.77), respectively; the PPV was 0.13 with sensitivity at 0.81 at the PACU time point. The area under the precision-recall curves (AUPRC) were essentially the same with the 3 models; the AUPRCs for the LASSO model at anesthesia induction, end of surgery, and PACU time points were 0.18 (95% CI: 0.16-0.21), 0.19 (95%CI: 0.17-0.22), and 0.20 (95%CI: 0.18-0.23), respectively. The calibration curves for the models and the empirical cumulative density function graphs are presented in [Supplementary Digital Content Figure S3](#), while the Brier skill scores are available in [Supplementary Digital Content Table S5](#).

The results of the univariate and multivariable analysis are presented in [Supplementary Digital Content Table S6](#). Decreasing age was the only demographic variable positively associated with emergence delirium. None of the patient class or past medical history variables were associated with emergence delirium. Adenoidectomy with or without tonsillectomy procedures were most strongly associated with emergence delirium occurrence, followed by ophthalmology

and ear tube surgery. Minor general surgery/urology procedures had a negative association with emergence delirium. A higher incidence of delirium was associated with surgery being performed at an ambulatory surgery center, absence of a preoperative blood pressure measurement, and in patients who were premedicated with midazolam. Intraoperative ondansetron was positively associated with emergence delirium while ketorolac had a negative association, as did the patient being asleep or unconscious at the first assessment following PACU arrival.

[Supplementary Digital Content Table S7](#) shows parsimonious models' performance in the secondary analysis. All the parsimonious models performed lower than their counterpart full-size models with statistical significance.

Discussion

We present the findings of applying 4 ML algorithms on a large quaternary hospital EHR dataset to develop a predictive model for pediatric emergence delirium. There were little differences in the performance of the models studied, with

Table 1. Patient characteristics with odds ratios in 2 outcome groups (emergence delirium vs none).

Variable	Emergence delirium		Odds ratio (95% CI)
	No [<i>n</i> = 50 420] (%)	Yes [<i>n</i> = 4356] (%)	
Demographics			
Sex			
Female	20 990 (41.63)	1700 (39.03)	3.13 (2.15-4.67)
Male	29 430 (58.37)	2656 (60.97)	reference
Race			
White	30 591 (61.42)	2692 (62.91)	reference
Black	8740 (17.55)	727 (16.99)	1.67 (1.17-2.37)
Other ^b	10 474 (14.2)	937 (21.51)	0.68 (0.35-1.2)
Unknown	615 (0.01)	0 (0.00)	N/A
Hispanic or Latino			
Yes	4798 (9.52)	405 (9.30)	1.26 (0.64-2.23)
No	45 622 (90.48)	3951 (90.70)	reference
Unknown	0 (0.0)	0 (0)	N/A
Age at surgery, years ^a	—	—	0.82 (0.80-0.83)
Former prematurity ^a	—	—	0.95 (0.90-1.00)
Body mass index ^a	—	—	0.95 (0.94-0.96)
Past medical history			
ADHD/ADD	2910 (5.77)	187 (4.29)	0.73 (0.63-0.85)
Asthma/reactive airway disease	10 612 (21.05)	941 (21.60)	1.03 (0.96-1.11)
Autism/autism spectrum disorder	2516 (4.99)	227 (5.21)	1.05 (0.91-1.20)
Developmental delay	7668 (15.21)	698 (16.02)	1.06 (0.98-1.16)
Obstructive sleep apnea	6554 (13.00)	917 (21.05)	1.78 (1.65-1.93)
Seizure disorder	2,335 (4.63)	162 (3.72)	0.80 (0.68-0.94)
History of prior surgery	30 878 (61.24)	2407 (55.26)	0.78 (0.73-0.83)
Patient class			
Day surgery	37 155 (73.69)	3412 (78.33)	reference
Admit before surgery	130 (0.26)	2 (0.05)	0.17 (0.04-0.68)
Admit after surgery (inpatient)	2243 (4.45)	115 (2.64)	0.56 (0.46-0.68)
Admit after surgery (observation)	5185 (10.28)	663 (15.22)	1.39 (1.27-1.52)
Emergency room	1233 (2.45)	24 (0.55)	0.21 (0.14-0.32)
Inpatient	4474 (8.87)	140 (3.21)	0.34 (0.29-0.40)
Premedication			
Midazolam	40 841 (81.00)	3971 (91.16)	2.45 (2.25-2.73)
Dexmedetomidine	141 (0.28)	13 (0.30)	2.25 (1.24-4.10)
Diazepam	307 (0.61)	9 (0.21)	0.90 (0.44-1.84)
Ketamine	116 (0.23)	2 (0.05)	0.63 (0.15-2.57)
More than 1 premedication documented	114 (0.23)	2 (0.05)	0.44 (0.11-1.79)
No premedication documented	9130 (18.11)	363 (8.33)	reference

Abbreviations: ADHD = attention deficit hyperactivity disorder; ADD = attention deficit disorder.

^a Continuous variable.

^b Includes Asian, Indian, Native American Indian, Alaska Islander, Native Hawaiian, and other Pacific islander.

AUROC ranging between 0.74 and 0.75. At a sensitivity level of 0.8, specificity for the 4 models ranged from 0.5 to 0.57, and PPV ranged from 0.13 to 0.14. AUPRCs, representing expected PPVs across different sensitivities, ranged between 0.18 and 0.2 across the models and 3 time points, which are at least 2 times higher than the emergence delirium prevalence of 0.08.

We recognized that the selection of operating points (predicted probability thresholds) for each model leads to varying underlying performance characteristics. The prediction thresholds we selected can favor sensitivity, that is, with a value of 0.57 specificity and 0.8 sensitivity at the PACU arrival time point, we will be able to capture 80% of events at the cost of a 57% false-positive rate, and the corresponding PPV was 0.14. Although a downside of this approach is that false-positives might receive treatment to prevent delirium, the 0.14 PPV based on the high-sensitivity threshold is higher than the cohort outcome prevalence of 0.08 (a 75% increase). It makes sense to favor higher sensitivity based on the following points: (1) a higher percentage of patients at

risk can be covered, especially within this population with a very low emergence delirium prevalence (0.08), (2) the PPV of the corresponding operation threshold is higher than the emergence delirium prevalence, and (3) clinicians would take the risks of preventative therapies into consideration when making treatment decisions.

The ML model we developed underperformed relative to the Emergence Agitation Risk Scale (AUROC 0.81 95% CI 0.72-0.89), which was the only predictive model for pediatric emergence delirium that met methodologic criteria in a recent systematic review.¹⁶ However, this scale's validation cohort was only 100 patients and included younger patients undergoing a much more limited set of procedures.¹⁵ Our modeling results could serve as a first step in developing clinical decision support, but further optimization and validation are needed. Although the XGB model had slightly improved performance, we believe this is offset by its interpretability. We favor using the LASSO model as its underlying analysis and variables are more transparent to clinicians compared to "black-box" models like RF and XGB.

Table 2. List of tunable hyperparameters of machine learning models and results of model testing.

Model and parameter range	Selected parameters	AUROC		AUPRC		Sensitivity		Specificity		Positive predictive value		Negative predictive value		F1-score	
		Training (95% CI)	Testing (95% CI)	Training (95% CI)	Testing (95% CI)	Training (95% CI)	Testing (95% CI)	Training (95% CI)	Testing (95% CI)	Training (95% CI)	Testing (95% CI)	Training (95% CI)	Testing (95% CI)	Training (95% CI)	Testing (95% CI)
LASSO															
C = 0.1															
Anesthesia induction		0.72 (0.72-0.73)	0.72 (0.70-0.75)	0.18 (0.18-0.19)	0.18 (0.16-0.21)	0.80 (0.79-0.81)	0.80 (0.76-0.83)	0.51 (0.50-0.51)	0.51 (0.49-0.52)	0.12 (0.12-0.12)	0.12 (0.12-0.13)	0.97 (0.97-0.97)	0.97 (0.96-0.97)	0.21 (0.21-0.22)	0.21 (0.20-0.22)
End of surgery		0.73 (0.72-0.74)	0.73 (0.70-0.75)	0.19 (0.18-0.20)	0.19 (0.17-0.22)	0.81 (0.80-0.82)	0.80 (0.77-0.84)	0.52 (0.51-0.52)	0.52 (0.50-0.53)	0.13 (0.12-0.13)	0.13 (0.12-0.13)	0.97 (0.97-0.97)	0.97 (0.96-0.97)	0.22 (0.21-0.22)	0.22 (0.21-0.23)
Through PACU		0.74 (0.74-0.75)	0.74 (0.72-0.77)	0.20 (0.19-0.21)	0.20 (0.18-0.23)	0.81 (0.80-0.82)	0.81 (0.77-0.84)	0.54 (0.53-0.54)	0.54 (0.52-0.55)	0.13 (0.13-0.13)	0.13 (0.12-0.14)	0.97 (0.97-0.97)	0.97 (0.96-0.98)	0.23 (0.22-0.23)	0.22 (0.21-0.23)
Random Forest															
Number of trees = 150															
Anesthesia induction	Max depth = 9	0.82 (0.81-0.83)	0.72 (0.70-0.75)	0.38 (0.36-0.39)	0.18 (0.16-0.21)	0.87 (0.86-0.88)	0.78 (0.74-0.81)	0.57 (0.57-0.57)	0.56 (0.54-0.57)	0.15 (0.15-0.15)	0.13 (0.12-0.14)	0.98 (0.98-0.98)	0.97 (0.96-0.97)	0.25 (0.25-0.26)	0.22 (0.21-0.23)
End of surgery	Max depth = 12	0.85 (0.84-0.85)	0.73 (0.71-0.75)	0.40 (0.39-0.42)	0.19 (0.17-0.22)	0.94 (0.93-0.94)	0.78 (0.74-0.82)	0.51 (0.51-0.52)	0.57 (0.56-0.58)	0.14 (0.14-0.14)	0.14 (0.13-0.14)	0.99 (0.98-0.99)	0.97 (0.96-0.97)	0.25 (0.24-0.25)	0.23 (0.22-0.24)
Through PACU	Max depth = 12	0.88 (0.87-0.88)	0.74 (0.72-0.77)	0.45 (0.44-0.47)	0.20 (0.18-0.23)	0.96 (0.96-0.97)	0.84 (0.80-0.87)	0.53 (0.53-0.54)	0.50 (0.50-0.52)	0.15 (0.15-0.15)	0.13 (0.12-0.13)	0.99 (0.99-1.00)	0.97 (0.97-0.98)	0.26 (0.26-0.26)	0.22 (0.21-0.23)
Ridge															
C = 0.1															
Anesthesia induction		0.73 (0.72-0.73)	0.72 (0.70-0.75)	0.18 (0.18-0.19)	0.18 (0.16-0.21)	0.80 (0.79-0.81)	0.80 (0.76-0.83)	0.51 (0.50-0.51)	0.51 (0.49-0.52)	0.12 (0.12-0.12)	0.12 (0.12-0.13)	0.97 (0.97-0.97)	0.97 (0.96-0.97)	0.21 (0.21-0.22)	0.21 (0.20-0.22)
End of surgery		0.73 (0.73-0.74)	0.73 (0.70-0.75)	0.19 (0.18-0.20)	0.19 (0.17-0.22)	0.81 (0.80-0.82)	0.80 (0.77-0.84)	0.52 (0.51-0.52)	0.52 (0.51-0.53)	0.13 (0.12-0.13)	0.13 (0.12-0.13)	0.97 (0.97-0.97)	0.97 (0.96-0.97)	0.22 (0.22-0.22)	0.22 (0.21-0.23)
Through PACU		0.75 (0.74-0.75)	0.74 (0.72-0.76)	0.20 (0.19-0.21)	0.20 (0.17-0.23)	0.81 (0.80-0.82)	0.80 (0.77-0.84)	0.54 (0.53-0.54)	0.54 (0.53-0.55)	0.13 (0.13-0.13)	0.13 (0.12-0.14)	0.97 (0.97-0.97)	0.97 (0.96-0.97)	0.23 (0.22-0.23)	0.22 (0.22-0.23)
XGB															
Number of trees = 100															
Anesthesia induction	Max depth = 3	0.75 (0.74-0.76)	0.72 (0.70-0.75)	0.22 (0.21-0.23)	0.18 (0.16-0.21)	0.82 (0.81-0.83)	0.79 (0.76-0.83)	0.53 (0.52-0.53)	0.53 (0.51-0.54)	0.13 (0.13-0.13)	0.13 (0.12-0.13)	0.97 (0.97-0.97)	0.97 (0.96-0.97)	0.22 (0.22-0.23)	0.22 (0.21-0.23)
End of surgery	Max depth = 6	0.77 (0.77-0.78)	0.73 (0.71-0.76)	0.24 (0.23-0.26)	0.19 (0.17-0.22)	0.84 (0.83-0.85)	0.80 (0.76-0.83)	0.55 (0.55-0.55)	0.55 (0.53-0.56)	0.14 (0.14-0.14)	0.13 (0.13-0.14)	0.97 (0.97-0.98)	0.97 (0.96-0.97)	0.24 (0.23-0.24)	0.23 (0.22-0.24)
Through PACU	Max depth = 6	0.80 (0.79-0.80)	0.75 (0.72-0.77)	0.27 (0.26-0.29)	0.20 (0.18-0.24)	0.86 (0.85-0.87)	0.80 (0.77-0.84)	0.57 (0.57-0.58)	0.57 (0.55-0.58)	0.15 (0.15-0.15)	0.14 (0.13-0.14)	0.98 (0.98-0.98)	0.97 (0.96-0.98)	0.25 (0.25-0.26)	0.24 (0.23-0.25)

Abbreviations: AUROC = area under the receiver operating characteristic curve; AUPRC = area under the precision-recall curve; PACU = post anesthesia care unit.

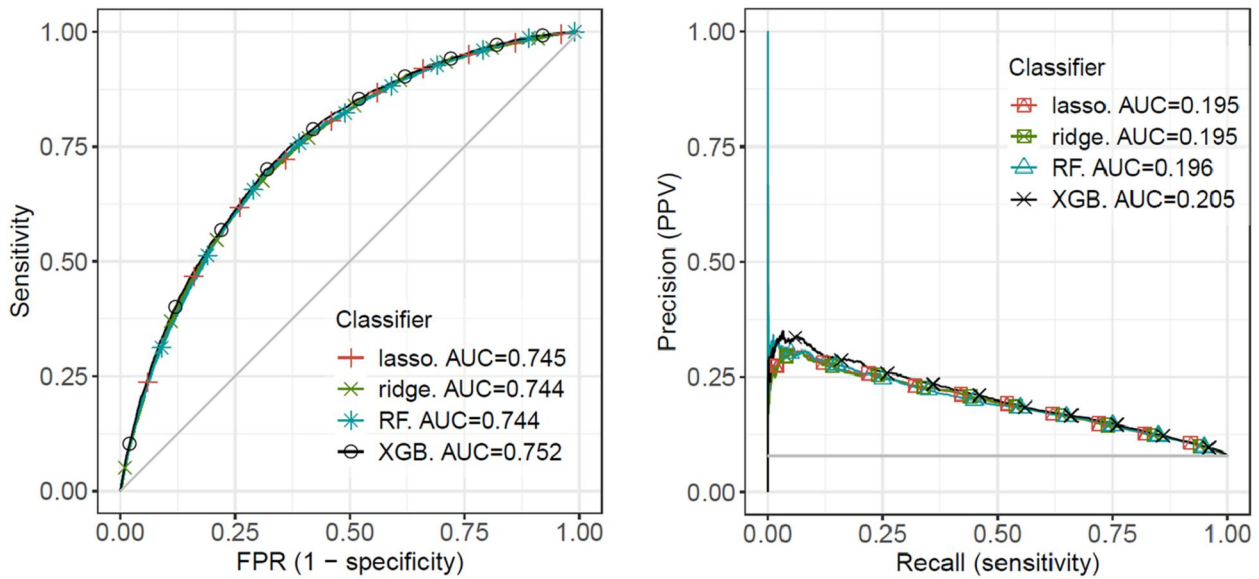


Figure 2. Receiver operating characteristic curves and precision versus sensitivity plots for the 4 studied machine learning models. Abbreviation: FPR = false positive rate; PPV = positive predictive value; RF = random forest; XGB = Extreme Gradient Boost.

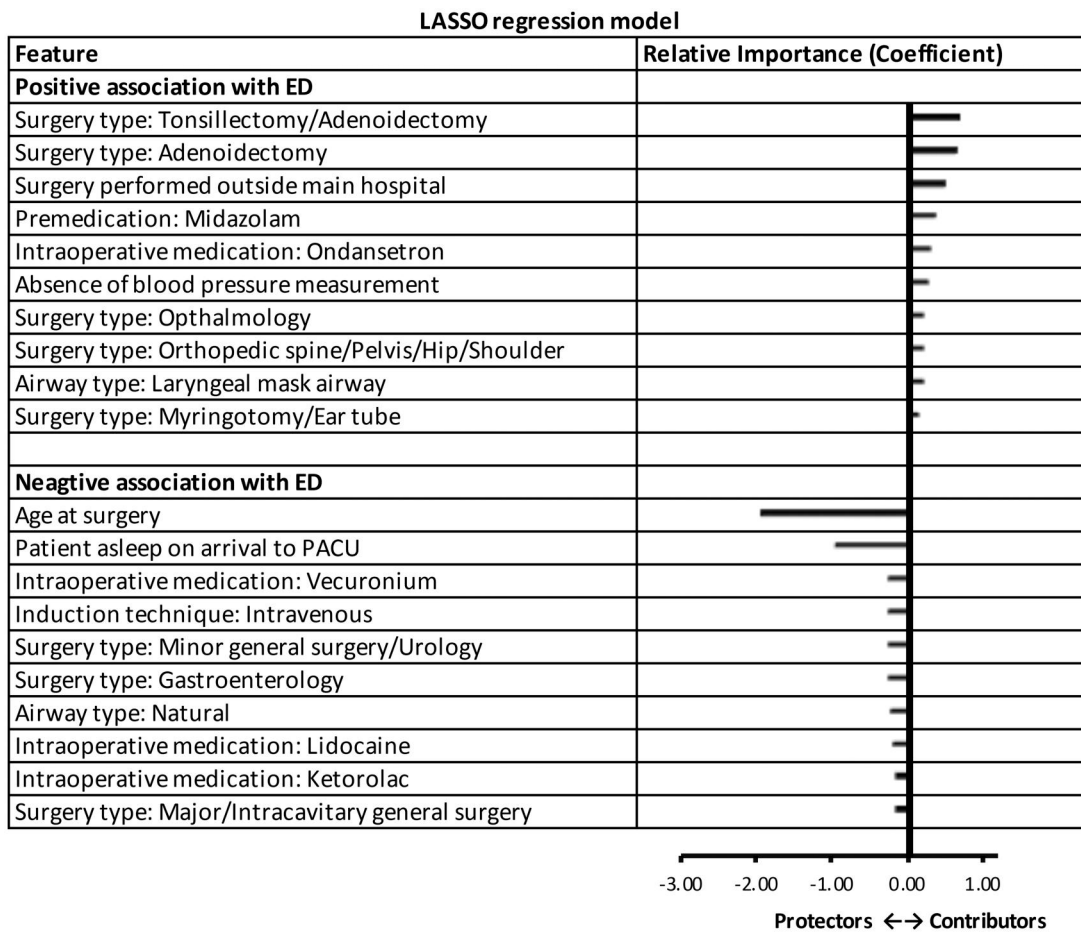


Figure 3. Top 10 variables associated with increased and decreased likelihood of emergence delirium, ordered based on LASSO regression model coefficient.

We discovered an association of midazolam premedication with emergence delirium, both in the LASSO model and multivariable analysis. Several studies have evaluated midazolam for preventing emergence delirium,^{32,33} and it has been

suggested as treatment.³ Yet, benzodiazepines have been implicated in promoting intensive care unit (ICU) delirium both in adults^{34,35} and in children.^{36,37} Our finding parallels these studies, although the reasons for this observation are

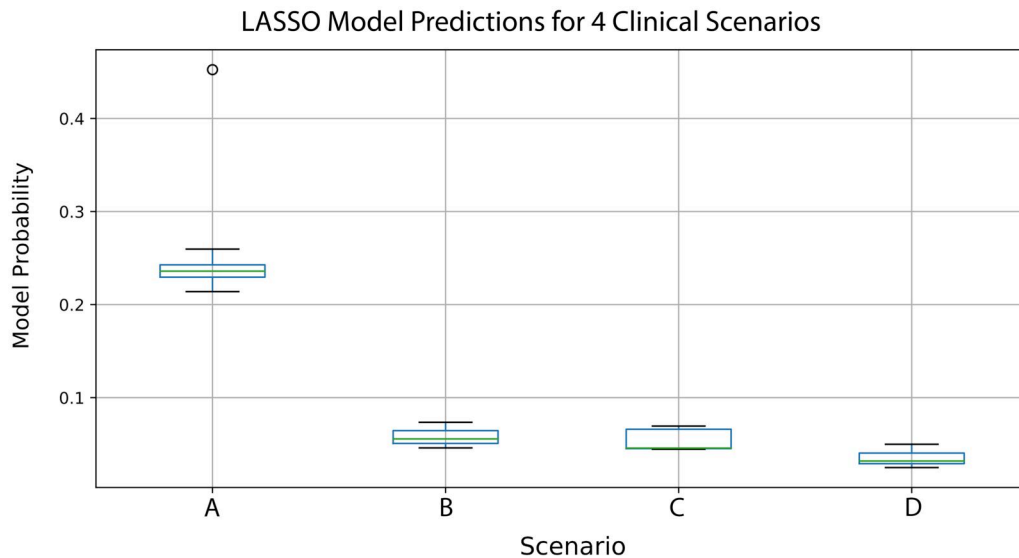


Figure 4. LASSO model predictions for 4 clinical scenarios. (Scenario A) A 3-year-old male, tonsillectomy/adenoidectomy, admit after surgery observation, midazolam premedication, preoperative blood pressure absent, inhaled anesthesia induction, moderate induction behavior. (Scenario B) A 5-year-old female, minor general surgery/urology procedure, day surgery, midazolam premedication, preoperative blood pressure present, inhaled anesthesia induction, moderate induction behavior. (Scenario C) A 5-year-old male, minor general surgery/urology procedure, day surgery, no premedication, preoperative blood pressure present, inhaled anesthesia induction, calm induction behavior. (Scenario D) A 9-year-old male, orthopedic extremity procedure, day surgery, midazolam premedication, preoperative blood pressure present, inhaled anesthesia induction, calm induction behavior.

unclear. Clinicians may consider this when deciding whether to administer midazolam premedication, particularly in patients with a history of emergence delirium. In our population, midazolam premedication is the default practice in the perioperative setting. The absence of midazolam premedication in our cohort may be a marker for temperament, lower anxiety, stronger coping skills, etc., and this may be a confounder for our finding of an association of midazolam with emergence delirium.

Some study findings are curious. Ondansetron was positively associated emergence delirium in both the LASSO model and multivariable analysis. If replicated in further studies, this information could be relevant to decisions around anti-emetic prophylaxis. An intravenous induction was protective in the LASSO model, while anesthesia maintenance technique was not, suggesting the possibility that high sevoflurane concentrations with inhalation induction may play a role in emergence delirium. One systematic review by Haque et al³⁸ reported that ondansetron may be an effective agent for preventing or treating postoperative delirium in adult populations; however, the authors also stated that such conclusions from the studies are tenuous given they are few and of poor quality. Pediatric emergence delirium, on the other hand, is a distinct clinical entity from postoperative delirium in adults. Pediatric emergence delirium occurs in the immediate period following emergence from anesthesia, whereas postoperative delirium can develop days following surgery. Our findings for a positive association between ondansetron and emergence delirium may generate further evaluation inquiries.

There was little decrement in model performance when it only included data that were available at the time of anesthesia induction. This could be useful in that most of the model's predictive power is available at the start of surgery, and consequently evidence-based interventions may be employed to reduce the likelihood of emergence delirium. However, some such interventions had negligible coefficients in the LASSO

model and were not significant in the multivariable analysis. These interventions included anesthesia maintenance with propofol, bolus propofol at the end of the case, and intraoperative dexmedetomidine. In this large and heterogeneous patient sample, effects of these interventions may not be as readily identified as they might be in controlled trials with homogeneously selected and managed sample populations.

We observed a strong association with increasing age and decreased incidence of emergence delirium. This is consistent with other studies^{3,5,9}; however, the biological underpinnings of this phenomenon remain unexplained. Also consistent with at least one prior study is our finding of an association with adenoidectomy with or without tonsillectomy.³ It may be that both degree and location (eg, throat, eye) of surgical tissue trauma impact the likelihood of emergence delirium. Ketorolac was also negatively associated with emergence delirium, an association reported in the setting of tympanostomy tube insertion.⁷ Ketorolac's observed effect may be mediated through pain and inflammation pathways.

Limitations

This retrospective study has limitations. While the findings we report are objective and from a reproducible study design, the interpretations may be subject to hindsight bias and narrative fallacy. The primary outcome measure, the Watcha scale, has not been as extensively validated as the pediatric anesthesia emergence delirium (PAED) scale,⁹ and as with other delirium assessments there may be overlap with pain behaviors. Nevertheless, a crying inconsolable child (Watcha = 3) or a child who is agitated and thrashing (Watcha = 4) represent an undesirable outcome. Supporting our use of this measure, the Watcha scale has been shown to perform comparably to the PAED scale, is more readily implemented in routine clinical practice, and may be more sensitive and specific.²⁴ Additionally, the procedure and maintenance of anesthesia categorizations we used were imperfect. While imperfect, the schemas employed are reproducible.

Several of the studied variables covary. Examples include younger age and absence of a preoperative blood pressure measurement, airway type and surgical procedure, and lidocaine administration and intravenous induction. We deliberately included a large number of variables because a purported strength of ML algorithms is that they effectively deal with large numbers of inter-related variables. Our use of encounter-level data rather than patient-level data is another possible limitation; our rationale for this is described above, and patients with a single encounter comprised 85% of the study sample.

Another limitation is that the management and outcome data are confounded to some extent and likely reflect treating anesthesiologists' knowledge and practices directed at avoiding emergence delirium based on patient condition and clinical circumstance. Lastly, the single institution data used likely reflects local biases and practices, which may reduce the generalizability of the results. However, predictive models may perform optimally when rooted in data from the health system in which they are deployed.

Conclusions

In this study of emergence delirium using a large pediatric EHR dataset, ML modeling achieved an area under the ROC curve of 0.74-0.75 and an area under the PRC of 0.18-0.2 compared to the incidence of 0.08. Some study findings warrant further investigation, such as the association of midazolam premedication and intraoperative ondansetron administration with emergence delirium. Future actions will be to evaluate whether clinical decision making guided by the developed model yields improvements in patient outcomes.

Author contributions

H.Y.: This author helped design the study, conduct the study, analyze the data, has seen the original study data, dataset preparation, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript. A.F.S.: This author helped design the study, conduct the study, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript. V.M.R.: This author helped conduct the study, analyze the data, has seen the original study data, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript. O.N.: This author helped design the study, conduct the study, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript. W.T.M.: This author helped design the study, conduct the study, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript. T.N.S.: This author helped design the study, conduct the study, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript. J.A.G.: This author helped design the study, conduct the study, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript. M.B.P.: This author helped design the study, conduct the study, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript. P.A.S.: This author helped design the study, conduct the study, analyze the data, has seen the original study data, dataset preparation, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript. F.T.: This author helped design the study, conduct the

study, analyze the data, has seen the original study data, dataset preparation, reviewed the analysis of the data, write the manuscript, and has approved the final manuscript.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This research was supported by the Children's Hospital of Philadelphia.

Conflicts of interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced or appeared to influence the work reported in this paper.

Data availability

The data underlying this article were extracted from the electronic health record at the study site and cannot be shared publicly for the privacy of individuals who participated in the study.

References

1. Przybylo HJ, Martini DR, Mazurek AJ, Bracey E, Johnsen L, Coté CJ. Assessing behaviour in children emerging from anaesthesia: can we apply psychiatric diagnostic techniques? *Paediatr Anaesth*. 2003;13(7):609-616. <https://doi.org/10.1046/j.1460-9592.2003.01099.x>
2. Eckenhoff JE, Kneale DH, Dripps RD. The incidence and etiology of postanesthetic excitement. A clinical survey. *Anesthesiology*. 1961;22:667-673. <https://doi.org/10.1097/00000542-196109000-00002>
3. Voepel-Lewis T, Malviya S, Tait AR. A prospective cohort study of emergence agitation in the pediatric postanesthesia care unit. *Anesth Analg*. 2003;96(6):1625-1630. <https://doi.org/10.1213/01.ane.0000062522.21048.61>
4. Moore JK, Moore EW, Elliott RA, St Leger AS, Payne K, Kerr J. Propofol and halothane versus sevoflurane in paediatric day-case surgery: induction and recovery characteristics. *Br J Anaesth*. 2003;90(4):461-466. <https://doi.org/10.1093/bja/aeg098>
5. Aono J, Ueda W, Mamiya K, Takimoto E, Manabe M. Greater incidence of delirium during recovery from sevoflurane anesthesia in preschool boys. *Anesthesiology*. 1997;87(6):1298-1300. <https://doi.org/10.1097/00000542-199712000-00006>
6. Cole JW, Murray DJ, McAllister JD, Hirshberg GE. Emergence behaviour in children: defining the incidence of excitement and agitation following anaesthesia. *Paediatr Anaesth*. 2002;12(5):442-447. <https://doi.org/10.1046/j.1460-9592.2002.00868.x>
7. Davis PJ, Greenberg JA, Gendelman M, Fertal K. Recovery characteristics of sevoflurane and halothane in preschool-aged children undergoing bilateral myringotomy and pressure equalization tube insertion. *Anesth Analg*. 1999;88(1):34-38.
8. Murray DJ, Cole JW, Shrock CD, Snider RJ, Martini JA. Sevoflurane versus halothane: effect of oxycodone premedication on emergence behaviour in children. *Paediatr Anaesth*. 2002;12(4):308-312. <https://doi.org/10.1046/j.1460-9592.2002.00789.x>
9. Sikich N, Lerman J. Development and psychometric evaluation of the pediatric anesthesia emergence delirium scale. *Anesthesiology*. 2004;100(5):1138-1145. <https://doi.org/10.1097/00000542-200405000-00015>

10. Jerome EH. Recovery of the pediatric patient from anesthesia. In: Gregory GA, ed. *Pediatric Anesthesia*. 2nd ed. New York: Churchill Livingstone; 1989:629.
11. Malarbi S, Stargatt R, Howard K, Davidson A. Characterizing the behavior of children emerging with delirium from general anesthesia. *Paediatr Anaesth*. 2011;21(9):942-950. <https://doi.org/10.1111/j.1460-9592.2011.03646.x>
12. Olympio MA. Postanesthetic delirium: historical perspectives. *J Clin Anesth*. 1991;3(1):60-63. [https://doi.org/10.1016/0952-8180\(91\)90209-6](https://doi.org/10.1016/0952-8180(91)90209-6)
13. Kain ZN, Caldwell-Andrews AA, Maranets I, et al. Preoperative anxiety and emergence delirium and postoperative maladaptive behaviors. *Anesth Analg*. 2004;99(6):1648-1654. <https://doi.org/10.1213/01.ANE.0000136471.36680.97>
14. Kim J, Byun SH, Kim JW, et al. Behavioral changes after hospital discharge in preschool children experiencing emergence delirium after general anesthesia: a prospective observational study. *Paediatr Anaesth*. 2021;31(10):1056-1064. <https://doi.org/10.1111/pan.14259>
15. Hino M, Mihara T, Miyazaki S, et al. Development and validation of a risk scale for emergence agitation after general anesthesia in children: a prospective observational study. *Anesth Analg*. 2017;125(2):550-555. <https://doi.org/10.1213/ANE.00000000000002126>
16. Petre MA, Saha B, Kasuya S, et al. Risk prediction models for emergence delirium in paediatric general anaesthesia: a systematic review. *BMJ Open*. 2021;11(1):e043968. <https://doi.org/10.1136/bmjopen-2020-043968>
17. Georgiyants M A, Pushkar M B, Vysotska E V, Porvan A P. Optimization of current postoperative period after childrens' adenotomy. *Lik Sprava*. 2017;1-2:115-119. [https://doi.org/10.31640/LS-2017\(1-2\)18](https://doi.org/10.31640/LS-2017(1-2)18)
18. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318. <https://doi.org/10.1001/jama.2017.18391>
19. Bishara A, Chiu C, Whitlock EL, et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. *BMC Anesthesiol*. 2022;22(1):8. <https://doi.org/10.1186/s12871-021-01543-y>
20. Neto PCS, Rodrigues AL, Stahlschmidt A, Helal L, Stefani LC. Developing and validating a machine learning ensemble model to predict postoperative delirium in a cohort of high-risk surgical patients: a secondary cohort analysis. *Eur J Anaesthesiol*. 2023;40(5):356-364. <https://doi.org/10.1097/EJA.0000000000001811>
21. Song YX, Yang XD, Luo YG, et al. Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: a retrospective study. *CNS Neurosci Ther*. 2023;29(1):158-167. <https://doi.org/10.1111/cns.13991>
22. Snell KIE, Levis B, Damen JAA, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;381:e073538. <https://doi.org/10.1136/bmj.g7594>
23. Watcha MF, Ramirez-Ruiz M, White PF, Jones MB, Laguieruela RG, Terkonda RP. Perioperative effects of oral ketorolac and acetaminophen in children undergoing bilateral myringotomy. *Can J Anaesth*. 1992;39(7):649-654. <https://doi.org/10.1007/BF03008224>
24. Bajwa SA, Costi D, Cyna AM. A comparison of emergence delirium scales following general anesthesia in children. *Paediatr Anaesth*. 2010;20(8):704-711. <https://doi.org/10.1111/j.1460-9592.2010.03328.x>
25. Uezono S, Goto T, Terui K, et al. Emergence agitation after sevoflurane versus propofol in pediatric patients. *Anesth Analg*. 2000;91(3):563-566. <https://doi.org/10.1097/00000539-200009000-00012>
26. Hmisc package reference. Secondary Hmisc package reference. Accessed September 11, 2023. <https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf>
27. Ruiz VM, Goldsmith MP, Shi L, et al. Early prediction of clinical deterioration using data-driven machine learning modeling of electronic health records. *J Thorac Cardiovasc Surg*. 2021;164(1):211-222.e3. <https://doi.org/10.1016/j.jtcvs.2021.10.060>
28. Shi L, Muthu N, Shaeffer GP, Sun Y, Ruiz Herrera VM, Tsui FR. Using data-driven machine learning to predict unplanned ICU transfers with critical deterioration from electronic health records. *Stud Health Technol Inform*. 2022;290:660-664. <https://doi.org/10.3233/SHTI220160>
29. Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F, eds. *Machine Learning and Knowledge Discovery in Databases*. ECML PKDD 2013. Lecture Notes in Computer Science(), Vol. 8190. Berlin, Heidelberg: Springer; 2013. https://doi.org/10.1007/978-3-642-40994-3_29
30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.
31. Fit.mult.impute reference. Secondary Fit.mult.impute reference. Accessed September 22, 2023. <https://rdrr.io/cran/Hmisc/src/R/fit.mult.impute.s>
32. Zhang C, Li J, Zhao D, Wang Y. Prophylactic midazolam and clonidine for emergence from agitation in children after emergence from sevoflurane anesthesia: a meta-analysis. *Clin Ther*. 2013;35(10):1622-1631. <https://doi.org/10.1016/j.clinthera.2013.08.016>
33. Lapin SL, Auden SM, Goldsmith LJ, Reynolds AM. Effects of sevoflurane anaesthesia on recovery in children: a comparison with halothane. *Paediatr Anaesth*. 1999;9(4):299-304. <https://doi.org/10.1046/j.1460-9592.1999.00351.x>
34. Riker RR, Shehabi Y, Bokesch PM, et al.; SEDCOM (Safety and Efficacy of Dexmedetomidine Compared With Midazolam) Study Group. Dexmedetomidine vs midazolam for sedation of critically ill patients: a randomized trial. *JAMA*. 2009;301(5):489-499. <https://doi.org/10.1001/jama.2009.56>
35. Zaal IJ, Devlin JW, Hazelbag M, et al. Benzodiazepine-associated delirium in critically ill adults. *Intensive Care Med*. 2015;41(12):2130-2137. <https://doi.org/10.1007/s00134-015-4063-z>
36. Traube C, Silver G, Gerber LM, et al. Delirium and mortality in critically ill children: epidemiology and outcomes of pediatric delirium. *Crit Care Med*. 2017;45(5):891-898. <https://doi.org/10.1097/CCM.0000000000002324>
37. Dervan LA, Di Gennaro JL, Farris RWD, Watson RS. Delirium in a tertiary PICU: risk factors and outcomes. *Pediatr Crit Care Med*. 2020;21(1):21-32. <https://doi.org/10.1097/PCC.00000000000002126>
38. Haque N, Naqvi RM, Dasgupta M. Efficacy of ondansetron in the prevention or treatment of post-operative delirium—a systematic review. *Can Geriatr J*. 2019;22(1):1-6. <https://doi.org/10.5770/cgj.22.266>