2021

A Machine Learning Approach to Predicting Early and Late Reintubation

Mathew J. Koretsky

A Machine Learning Approach to Predicting Early and Late Reintubation

Mathew Koretsky

Honors College Thesis

Advisor: Dr. Nick Cheney | Department of Computer Science

2020 – 2021

ABSTRACT

Accurate estimations of surgical risks is important for improving the shared decision making and informed consent processes. Reintubation is a severe postoperative complication that can lead to various other detrimental outcomes. Reintubation can also be broken up into early reintubation (within 72 hours of surgery) and late reintubation (within 30 days of surgery). Using clinical data provided by ACS NSQIP, scoring systems were developed for the prediction of combined, early, and late reintubation. The risk factors included in each scoring system were narrowed down from a set of 37 pre and perioperative factors. The scoring systems demonstrated good performance in terms of both accuracy and discrimination, and these results were only marginally worse than prediction using the full set of risk variables. While more work needs to be done to identify clinically relevant differences between the early and late reintubation outcomes, the scoring systems provided here can be used by surgeons and patients to improve the quality of care overall.

INTRODUCTION

Comprehending surgical risks is important for surgeons and patients throughout the shared decision-making process.[1] In the context of surgery, informed consent is the voluntary authorization of a surgical treatment by a patient with full comprehension of the risks involved.[2] Patients and surgeons both need information about surgical risks to discuss the possibility of alternative procedures and non-surgical treatments. Unfortunately, predicting postoperative risks has generally been based off of individual surgeon experience and does not take into account patient-specific risk factors.[3]

Identifying patient-specific risk factors associated with a range of surgical outcomes is an essential step in understanding and preventing these postoperative complications. While identifying these risk factors for a broad range of complications is important, efforts are often focused on more adverse outcomes such as mortality. One postoperative indicator for an increased risk of mortality is the need for unplanned reintubation, which is defined as requiring postoperative placement of an endotracheal tube in the 30 days following surgery.[4] In addition to this, unplanned reintubation is also associated with longer hospital stays and increased costs.[5,6] Therefore, determining risk factors that predispose patients to unplanned reintubation will help lead to fewer instances of detrimental complications.

Unplanned reintubation specifically is an intriguing complication to investigate as it can be broken up into two component outcomes. It is known that the risk of hypoxemia is greatest during the first 72 hours after surgery, and this is also the peak respiratory depressant period.[7] Reintubations that occur within this high-risk period, called early reintubation, are associated with a distinct set of risk factors and complications compared to reintubations performed after this period, which are referred to as late reintubation. It is important to thoroughly investigate the differences between these distinct postoperative events in order to fully understand the unfavorable unplanned reintubation outcome.

A range of factors that are independently associated with unplanned reintubation have been previously identified. However, none of these studies establish a widely accepted simple evaluation tool to assess the probability of unplanned reintubation occurring.[5] Moreover, these studies have been concerned with determining risks for early reintubation or the combined reintubation outcome.[4,5,7] These methodologies fail to draw comparisons between risk factors associated with early and late reintubation that could be clinically important. We were interested in determining the most important risk factors for predicting and distinguishing between postoperative early and late reintubation. Our goal is to use machine learning methods to identify the most important factors associated with the prediction of both early and late reintubation, and use these to develop and validate a simple scoring metric to estimate the probability of both outcomes occurring. This metric should be simple enough to use without requiring expensive computational power, and will be utilized by clinicians throughout the surgical decision-making process to maintain informed consent and improve patient outcomes.

METHODS

Data and Patients

Data for this study came from the database of the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP). The data were collected from over 500 participating hospitals in the U.S. for adult patients who underwent major surgical procedures across all subspecialties except for trauma and transplant.[8,9] For each case, a trained risk assessment nurse prospectively collected information for 273 variables including patient demographics, preoperative comorbidities, operative information, interoperative variables, and 30-day postoperative outcomes.[10] Further information regarding the ACS NSQIP data and its collection methods has been described extensively elsewhere.[11,12]

From the ACS NSQIP data, patients were identified who underwent surgical procedures between 2009 and 2017. As one would expect, the number of reintubation cases in the data (n=51,383) is very low compared to the total number of cases (n=6,002,268). This class imbalance can cause standard machine learning classification algorithms to ignore the minority class in the pursuit of accurate performance over the full range of instances and is a common problem that has been previously identified.[13] One way to alleviate this issue is to resample the minority instances, in our case patients who experience unplanned reintubation, until this class contains as many instances as the majority.[14] However, in this case since the reintubation class contains a relatively large number of instances, it is more sensible to simply randomly sample an equal number of instances from the non-reintubation cases to create a perfectly balanced dataset for prediction. This was done independently for the combined reintubation, early reintubation (n=19,422) and late reintubation (n=31,766) outcomes. It is important to note that for 195 of the reintubation cases in the data, the days until the reintubation variable was not available to create the distinction between early and late reintubation.

For the purposes of this study, patients with preoperative ventilator dependence and patients whose principal anesthesia techniques were not general anesthesia were removed from the analysis. Intubation of patients that are dependent on a ventilator prior to surgery is not considered unplanned, and it would not make sense to investigate reintubation for patients that were not initially intubated during the administration of general anesthesia. It is typical in this type of risk prediction analysis to exclude emergent and/or outpatient procedures, due to their high and low-acuity natures respectively.[3-5,7] Having an emergency surgery generally comes with increased risks and therefore could be a meaningful predictor for unplanned reintubation, which is why these cases are left in the analysis. Excluding outpatient cases could eliminate a substantial number of low-risk surgeries that prediction might still be meaningful for. After applying this exclusion criteria, the cohort for this analysis consisted of 92,731 cases for combined reintubation, 35,725 cases for early reintubation, and 56,679 cases for late reintubation in the NSQIP database.

Risk Factors

It is common in the literature to select a set of variables a priori based on predictive value for the primary outcome of interest, which in our case is unplanned reintubation. However, this can cause potentially important predictors to be left out of the analysis.[3,4] For the risk prediction models created here, any variable of clinical relevance that is included in the ACS NSQIP data and known pre or perioperatively is to be considered a candidate variable. This was done in order to capture any and all risk factors that may be associated with both early and late reintubation. This left 37 risk factors to be considered by the risk prediction models. Due to the categorical nature of many of these risk factors, some were broken down into multiple variables through techniques such as one-hot encoding to make the modelling process easier.[15] The 37 risk factors were encoded into 125 variables or features that were available to the precision models.

Demographic variables included age, race, and gender. Lifestyle variables included smoking (current smoker within one year of procedure). General factors included ASA classification, wound classification, transfer status, functional status, emergency status and body mass index. Preoperative comorbidities included history of severe COPD, ascites, congestive heart failure in 30 days before surgery, hypertension requiring medication, acute renal failure, currently on dialysis, disseminated cancer, open wound, steroid use for chronic condition, weight loss (>10% loss in body weight in the last 6 months), bleeding disorders, transfusion, diabetes, dyspnea, and sepsis in 48 hours before surgery. Laboratory or bloodwork variables were only considered if they were available for greater than 75% of cases included in the data. This left serum sodium, blood urea nitrogen, serum creatinine, white blood cell count, hematocrit, and platelet count. Finally, operative variables included elective status, surgical subspecialty, and present at the time of surgery (PATOS) variables for the following outcomes: surgical site infection, pneumonia, UTI, and sepsis. Missing data was handled using Buck's method for imputation, which is the standard ACS NSQIP modeling approach.[3,16] Extensive univariate analyses of these risk factors have been performed previously and are therefore left out of this analysis.[3,10,12]

To provide more operative information to the prediction models, a risk value was calculated for each current procedural terminology (CPT) code included in the data. ACS NSQIP has previously created procedure-specific risk prediction models that were used to generate CPT-specific risk values.[3] These risk values could then be used as a predictor in a generalized model that spans many different procedures. However, the procedure-specific models and CPT-specific risk values generated by ACS NSQIP are not publicly available. Therefore, a simple workaround had to be used in which for each CPT code, the risk value was assigned to be the proportion of surgeries where a postoperative complication occurred. This was done independently for each postoperative outcome. These CPT-specific risk values were calculated based on the full dataset and then mapped to the resampled datasets for each of the reintubation outcomes. This does induce a small amount of data leakage as bits of information from training sets will now be included in testing sets during modeling, however it should not affect the validity of the process overall. A histogram of the CPT-specific risk values for the combined reintubation, early reintubation and late reintubation outcomes are shown in Figure 1.

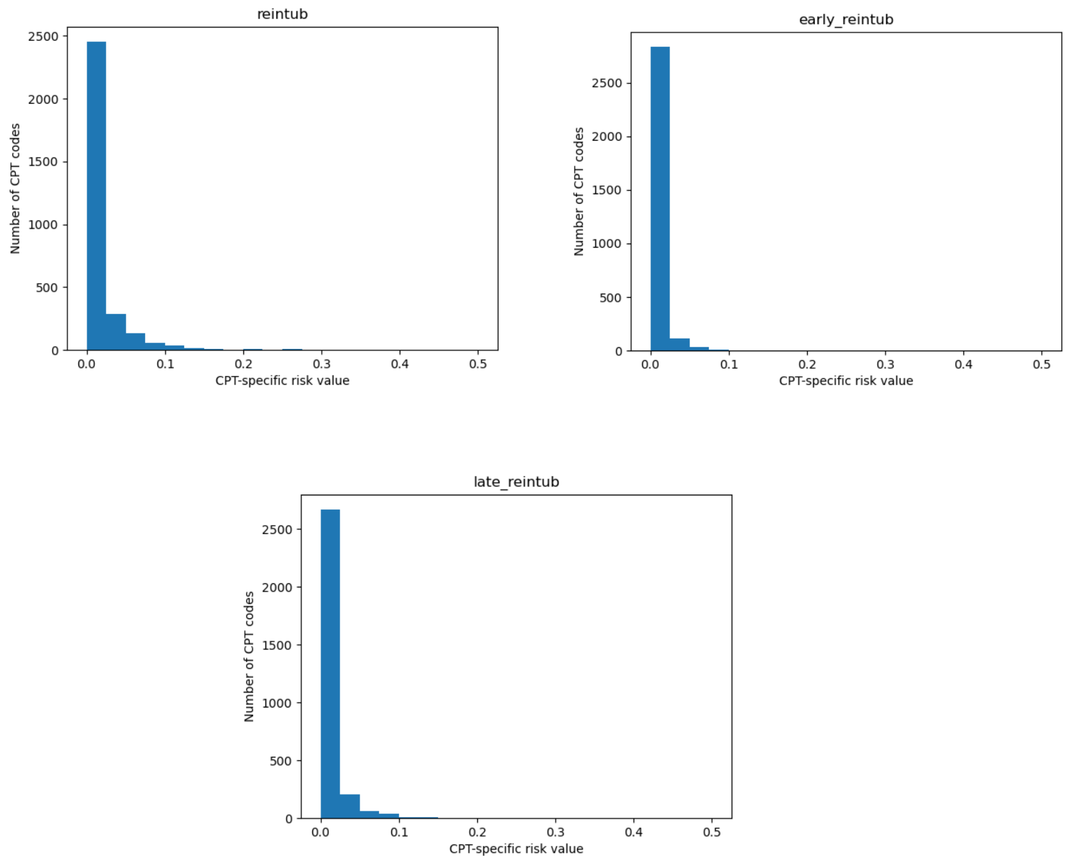

Figure 1: Histogram of CPT-specific risk values for combined reintubation (top left), early reintubation (top right) and late reintubation (bottom).

Statistical Analysis

In order to create a reasonable scoring metric for both early and late reintubation, the set of 125 encoded variables had to be narrowed down. Logistic regression, random forest classification, and gradient boosting classification models were initially fit on the full set of features using the created balanced datasets for combined reintubation, early reintubation and late reintubation. Logistic regression assumes underlying linear relationships in the data, hence the classification models were fit to assess different nonlinearities and determine which of the three had the best performance on the full set of features. Random forest and gradient boosting classification are both popular machine learning algorithms that use an ensemble of decision trees for prediction, and their methods have been published and discussed extensively elsewhere.[17,18] These three modeling approaches were also chosen because it is easy to extract the features that are important for prediction for each, through the coefficients from logistic regression and the importances from random forest and gradient boosting classification.

For each of the reintubation complications, the models were trained on 5 different 80% train/20% test splits. Cross-validation occurred on the first iteration to tune the values of the following hyperparameters: regularization strength and maximum iterations for logistic regression, and number of trees, maximum depth of trees and maximum features used per tree for the classification models. Each iteration, the models were evaluated on the test set using Brier score to assess probability prediction accuracy, and c-statistic to assess the discriminative power of the models. The Brier score is the average squared difference between patients' predicted probabilities and observed outcomes. The c-statistic is also known as the area under the receiver operating characteristic (ROC) curve, and ranges from 0.5 to 1, with a value of 0.5 indicating the model is doing no better than random at discriminating between events and non-events. The Brier score is the metric used for evaluation during cross-validation because it is computed from differences between the actual events and predicted probabilities, and is more informative than the c-statistic for our purposes. Another use for the Brier score is to get a benchmark value when the overall event rate in the training data is assigned to each patient, and this “null model” can be used to assess the added predictive contribution from the full set of features.

To determine which of the models demonstrated the best performance on the full set of features, a rank-sums test was utilized on the Brier score results from the 5 training and testing iterations for each of the three reintubation outcomes. This rank-based non-parametric test determines whether two samples come from distributions with the same median and is very well established in the field of statistics.[19] Two different methods were then used to extract the features that are most important for the prediction of each outcome. The first method is to use the covariates with either the top coefficients from the logistic regression models in terms of absolute value, or the top importances from the random forest and gradient boosting classification models. While these are not formally checked for significance, they still provide an idea of which variables are affecting the probability prediction most overall. The other method is to use the permutation importances evaluated on the test set that prove to be statistically significant at the 95% level. The permutation importance of a variable can be described as the gain or loss in some specified metric (i.e., Brier score, c-statistic) when that variable is randomly shuffled, or permuted, throughout the data. Detailed information regarding permutation importances and their methods have been published.[20,21]

One issue with this type of analysis is that the coefficients or importances, as well as the permutation importances, that show up as most important for the prediction of each outcome can be inconsistent between model runs. This is due to the large feature space being assessed and the specific train/test splits that are generated. Additionally, the number of features that show up as significant in terms of the permutation importances is dependent on the chosen evaluation metric. One other issue is that the CPT-specific risk values alone provide prediction that is similar in performance to using the full feature set, which can skew the predictive importance of other variables when these risk values are included.

To alleviate these issues and gain a better understanding of which features show up most often as important for prediction, a heuristic feature analysis was performed. Here, each of the models was run four times for each permutation importance evaluation metric being tested (i.e. Brier score and c-statistic), twice with the CPT-specific risk values included and twice without.

This led to 8 model runs for each of the algorithms and was performed for each of the three reintubation complications. The training and testing procedure stayed the same as before, with each run consisting of five iterations and cross-validation occurring on the first of those, tuning the same hyperparameters. From there, the number of times each feature showed up in the top 10 coefficients/importances and the significant permutation importances were tallied across the 8 model runs and subsequently ranked by their total appearances. The top 20 features in terms of total appearances identified by each model were isolated for each reintubation complication respectively.

The models were re-fit on this reduced set of features for each of the three complications. Cross-validation occurred on the set of 20 features, with the same hyperparameters being tuned as before. To establish the ideal number of risk factors to use in a scoring system, the set of 20 features was reduced iteratively by removing the least important in terms of total appearances, and refitting the models until only one feature (the most important) remained. On each iteration, the Brier score and c-statistic from the model was stored to allow for the creation of plots that show the performances vs. the number of features included. These will reveal the ideal number of features that should be used in a scoring system before performance falls off significantly, for each model and reintubation complication respectively. It is important to note that for this portion of the analysis, CPT-specific risk values are not included, because as discussed previously they can skew performance.

To establish a scoring system, the logistic regression models were once again refit, this time using the ideal number of features revealed from the plots. An 80% train/20% test split was utilized and cross-validation was once again performed to tune the regularization strength and maximum iteration hyperparameters. Here, the CPT-specific risk values were included to boost the performance. The coefficients from these refit models will be used to scale each variable, allowing for the creation of a reasonable scoring system for combined reintubation, early reintubation, and late reintubation respectively. While the heuristic feature analysis was performed for the classification models as well, the importances are not as interpretable in terms of scaling the covariates included in the system like the coefficients from the logistic regression models are.

RESULTS

From the ACS NSQIP database, three different sized cohorts of patients were identified for the analysis of combined reintubation (n=92,731), early reintubation (n=35,725), and late reintubation (n=56,679). These cases spanned all surgical subspecialties except for trauma and transplant, and consisted of 2,063 unique CPT codes being available for the creation of scoring systems for each reintubation complication.

Table 1 shows the results from fitting logistic regression, random forest classification, and gradient boosting classification models on the full set of features for each complication respectively. The values shown are the averages over the five training and testing iterations. Performance for late reintubation seems to be the best across the three models, both in terms of Brier score and c-statistic, however the values for the other complications are still reasonable.

	Logistic Regression		Random Forest		Gradient Boosting	
Outcome	Brier Score	C-stat	Brier score	C-stat	Brier score	C-stat
Reintubation	0.1401	0.8796	0.1439	0.8785	0.1328	0.8891
Early Reintubation	0.1598	0.8452	0.1514	0.8591	0.1456	0.8690
Late Reintubation	0.1359	0.8862	0.1304	0.8929	0.1226	0.9037

Table 1: Performance on the full set of features.

Tables 2, 3, and 4 contain the p-values from the rank sums tests based on the Brier score results from running the models on the full set of features for the combined reintubation, early reintubation, and late reintubation complications respectively. An asterisk indicates that the model in that row is performing significantly better than the model in that column. The null model scores were generated by assigning the overall event rate in the training data to each patient, and each of the models consistently outperforms this null model across all complications. Gradient boosting classification performs significantly better than both logistic regression and random forest classification across the reintubation outcomes. It is important to note that the p-values are very similar since comparing a distribution that only consists of five data points (for the five training and testing iterations) leads to a very limited number of comparisons that can be made.

Model	Null	Logistic Regression	Random Forest	Gradient Boosting
Logistic Regression	0.009*	x	0.047*	0.009
Random Forest	0.009*	0.047	x	0.009
Gradient Boosting	0.009*	0.009*	0.009*	x

Table 2: P-values from the rank-sums tests for the combined reintubation complication.

Model	null	Logistic Regression	Random Forest	Gradient Boosting
Logistic Regression	0.009*	x	0.009*	0.009
RF	0.009*	0.009	x	0.009
GBC	0.009*	0.009*	0.009*	x

Table 3: P-values from the rank-sums tests for the early reintubation complication.

Model	null	Logistic Regression	Random Forest	Gradient Boosting
Logistic Regression	0.009*	x	0.009*	0.009
RF	0.009*	0.009	x	0.009
GBC	0.009*	0.009*	0.009*	x

Table 4: P-values from the rank-sums tests for the late reintubation complication.

Table 5 displays the results from fitting the three models using only the CPT-specific risk values. While these results are statistically significantly worse than the results from fitting the models with the full feature set (Table 1), they are practically similar. The high level of performance when the CPT-specific risks are the sole predictor indicates that the importance of other variables can be skewed when these are included, and is the reason why half of the runs for the heuristic feature analysis do not contain these CPT-specific risks.

	Logistic Regression		Random Forest		Gradient Boosting	
Outcome	Brier Score	C-stat	Brier score	C-stat	Brier score	C-stat
Reintubation	0.1838	0.8315	0.1662	0.8312	0.1662	0.8312
Early Reintubation	0.2469	0.7980	0.1829	0.7979	0.1830	0.7971
Late Reintubation	0.1837	0.8558	0.1527	0.8556	0.1528	0.8552

Table 5: Performance on only the CPT-specific risk values.

Figure 2 displays the plots of performance vs. number of features that resulted from the logistic regression heuristic feature analysis for the three reintubation complications. For early reintubation, it is clear that the significant drop in performance occurs at 7 features. For both combined reintubation and late reintubation, there is a significant drop in performance around 12 and 15 features respectively, and another around 6 features. To keep the scoring system simplistic, the second significant drop in performance will be utilized.

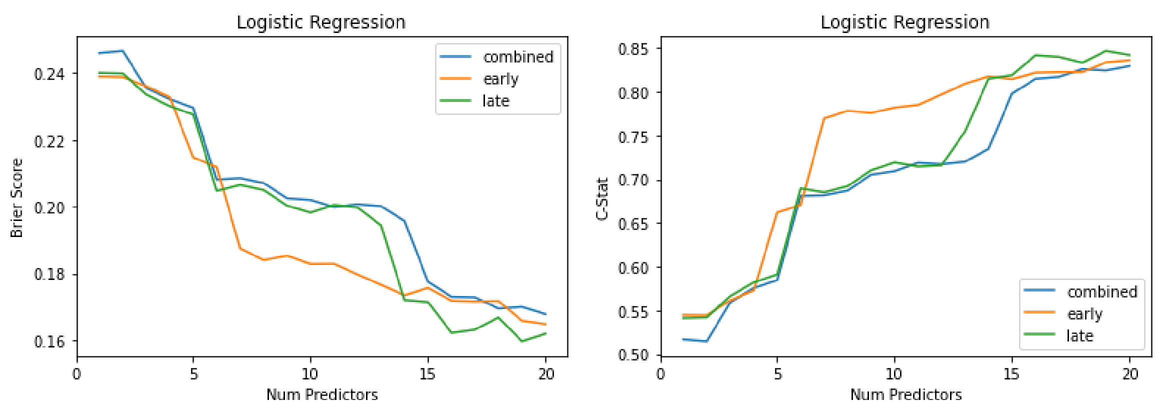

Figure 2: Results from the heuristic feature analysis for logistic regression across the three complications.

Figures 3 and 4 show the results from the heuristic feature analysis for random forest classification and gradient boosting classification. While these results are not used in the creation of a simple scoring system here, they could be incorporated into a more complex one that would require more computational power. The drop in performance for the random forest is seen around 9 features for combined reintubation, 6 features for early reintubation, and 8 features for late reintubation. For the gradient boosting models, the drop in performance comes at 6 features for combined reintubation and 8 features for early reintubation. Late reintubation never really sees a significant performance drop like the other two complications do, so it would be recommended to use a similar number of features, however more experimentation may be needed.

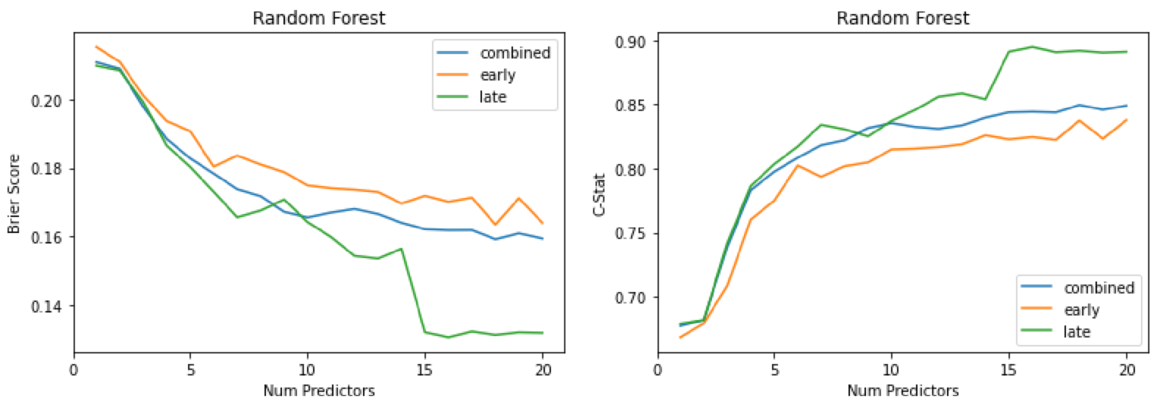

Figure 3: Results from the heuristic feature analysis for random forest across the three complications.

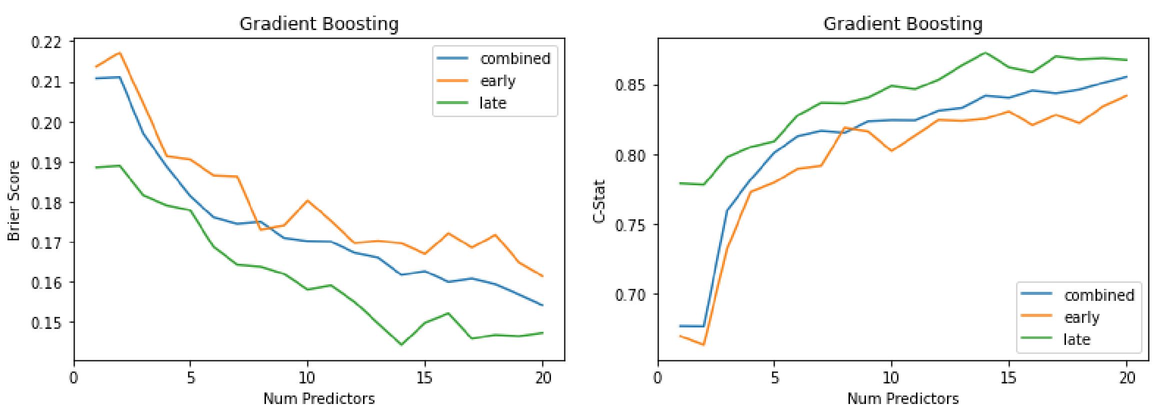

Figure 4: Results from the heuristic feature analysis for gradient boosting across the three complications.

Tables 6, 7 and 8 show the results from fitting the logistic regression models using the ideal number of features from the heuristic feature analysis (Figure 2). These coefficients will form the basis for a basic scoring system. The CPT risk values consistently carry the most weight in the prediction which is consistent from results seen from fitting models on the full set of features. There are some features that show up in the scoring system for all three reintubation complications, which is sensible because the outcomes only differ in timing. These include pneumonia present at the time of surgery (PATOS), septic shock PATOS, ASA 1: no disturb, and ASA 4: life threat. More interesting are the features that are distinct for each outcome, which include organ space surgical site infection (SSI) PATOS for combined reintubation, ascites and thoracic surgical subspecialty for early reintubation, and cardiac surgical subspecialty for late reintubation.

Variable	Coefficient
Reintubation CPT Risk	40.871
Pneumonia PATOS	2.590
ASA 1: No Disturb	-2.358
Septic Shock PATOS	2.044
ASA 5: Moribund	1.669
ASA 4: Life Threat	1.448
Organ Space SSI PATOS	0.848

Table 6: Combined reintubation scoring system.

Variable	Coefficient
Early Reintubation CPT Risk	9.034
ASA 1: No Disturb	-2.834
Septic Shock PATOS	2.374
Pneumonia PATOS	2.069
ASA 2: Mild Disturb	-1.608
ASA 4: Life Threat	1.216
Ascites	1.074
Surgical Subspecialty: Thoracic	0.914

Table 7: Early reintubation scoring system.

Variable	Coefficient
Late Reintubation CPT Risk	42.036
ASA 1: No Disturb	-2.715
Pneumonia PATOS	2.285
Septic Shock PATOS	2.227
ASA 4: Life Threat	1.581
ASA 5: Moribund	1.528
Surgical Subspecialty: Cardiac	0.214

Table 8: Late reintubation scoring system.

Finally, table 9 contains the Brier score and c-statistic results for running the logistic regression scoring systems. These values all lie in between the results seen when using the full feature set, and the results seen when CPT-specific risk values were used as the only predictor. Overall, the c-statistics show that for all three reintubation complications, the scoring systems are doing fairly well at discriminating between patients that have a reintubation and those who do not. The Brier scores show that the probability predictions are reasonable on average, with the scoring system for combined and late reintubation performing slightly better than that of early reintubation.

Outcome	Brier Score	C-stat
Reintubation	0.1674	0.8555
Early Reintubation	0.1825	0.8315
Late Reintubation	0.1668	0.8695

Table 9: Results from the logistic regression models using the scoring systems for the reintubation complications.

DISCUSSION

Accurate estimation of reintubation risks can help lead to less detrimental outcomes and better quality of patient care overall. Here, we set out to develop scoring systems for both early and late reintubation in order to determine which risk factors were most important for the prediction of each. These scoring systems have the potential to assist clinicians, patients and their families, as well as hospitals through its use.

Many existing reintubation risk assessment tools use a priori variable selection, which can leave potentially important factors out of these analyses, or only investigate the combined reintubation outcome. Here, we fit initial models on all available, relevant data (37 risk factors encoded as 125 variables), and demonstrate consistent performance across algorithms that assume underlying linear and nonlinear patterns in the data. These initial models also helped to determine the high level of performance that occurred when only CPT-specific risk values were used for prediction, and uncovered bias seen in the predictive importance of other included features.

From these initial models, a heuristic feature analysis was performed to determine which factors were most important for the prediction of combined, early, and late reintubation. Refitting these models on the reduced set of most important features allowed for the determination of the correct number to include in a scoring system. The scoring system displayed here is based on the logistic regression models, as the coefficients from these models are the most interpretable in terms of scaling each variable. The performance of the scoring systems for all three reintubation outcomes was reasonable in terms of both discrimination and accuracy. As expected, this performance lies in between the model performance when using the full set of features and the model performance when using only the CPT-specific risk values for prediction. There are some similarities in the features for the scoring systems determined for early and late reintubation, however there are also some important differences that can be further investigated. This work provides a foundation for discovering more differences between the distinct reintubation outcomes through tweaking the heuristic feature analysis process, and also looking at the important differences in features that result from the non-linear models.

There are some important limitations to the work that should be mentioned. First is that the ACS NSQIP data is only collected from approximately 10% of hospitals in the US.[3] The data is also collected from hospitals all over the country with variable capacity, and there are certainly variations in outcomes depending on the hospital and the surgeon performing an operation. The risk estimation tools presented here do not contain any information that may normalize this variation. Another limitation is that features that are tied in terms of importance from the heuristic feature analysis must still be ranked, and this is done in an arbitrary way. The ranking of feature importance is better determined by looking at the coefficients from the created scoring systems. Additional work must also be done to extract the clinical relevance from the differences seen in the scoring system and is better left to trained anesthesiology experts.

CONCLUSION

Accurate prediction of reintubation can lead to less instances of mortality after surgery, as well as lower costs and length of hospital stays. While there is still more work to be done in determining differentiating factors between early and late reintubation, the scoring systems created here can be used by patients and surgeons to promote discussions that can better inform the surgical decision making process. This in turn will improve patient understanding and informed consent, leading to better patient care overall.

REFERENCES

1. Knops, A. M., Legemate, D. A., Goossens, A., Bossuyt, P. M. M., & Ubbink, D. T. (2013). Decision Aids for Patients Facing a Surgical Treatment Decision: A Systematic Review and Meta-analysis. Annals of Surgery, 257(5), 860–866. https://doi.org/10.1097/SLA.0b013e3182864fd6

2. Leclercq, W. K. G., Keulers, B. J., Scheltinga, M. R. M., Spauwen, P. H. M., & van der Wilt, G.-J. (2010). A Review of Surgical Informed Consent: Past, Present, and Future. A Quest to Help Patients Make Better Decisions. World Journal of Surgery, 34(7), 1406–1415. https://doi.org/10.1007/s00268-010-0542-0

3. Bilimoria, K. Y., Liu, Y., Paruch, J. L., Zhou, L., Kmiecik, T. E., Ko, C. Y., & Cohen, M. E. (2013). Development and Evaluation of the Universal ACS NSQIP Surgical Risk Calculator: A Decision Aid and Informed Consent Tool for Patients and Surgeons. *Journal of the American College of Surgeons*, *217*(5), 833-842.e3. https://doi.org/10.1016/j.jamcollsurg.2013.07.385

4. Hua, M., Brady, J. E., & Li, G. (2012). A Scoring System to Predict Unplanned Intubation in Patients Having Undergone Major Surgical Procedures. *Anesthesia & Analgesia*, *115*(1), 88–94. https://doi.org/10.1213/ANE.0b013e318257012b

5. Brueckmann, B., Villa-Uribe, J. L., Bateman, B. T., Grosse-Sundrup, M., Hess, D. R., Schlett, C. L., & Eikermann, M. (2013). Development and Validation of a Score for Prediction of Postoperative Respiratory Complications. *Anesthesiology*, *118*(6), 1276–1285. https://doi.org/10.1097/ALN.0b013e318293065c

6. Beverly, A., Brovman, E. Y., Malapero, R. J., Lekowski, R. W., & Urman, R. D. (2016). Unplanned Reintubation Following Cardiac Surgery: Incidence, Timing, Risk Factors, and Outcomes. *Journal of Cardiothoracic and Vascular Anesthesia*, *30*(6), 1523–1529. https://doi.org/10.1053/j.jvca.2016.05.033

7. Ramachandran, S. K., Nafiu, O. O., Ghaferi, A., Tremper, K. K., Shanks, A., & Kheterpal, S. (2011). Independent Predictors and Outcomes of Unanticipated Early Postoperative Tracheal Intubation after Nonemergent, Noncardiac Surgery. *Anesthesiology*, *115*(1), 44–53. https://doi.org/10.1097/ALN.0b013e31821cf6de

8. Hall, B. L., Hamilton, B. H., Richards, K., Bilimoria, K. Y., Cohen, M. E., & Ko, C. Y. (2009). Does Surgical Quality Improve in the American College of Surgeons National Surgical Quality Improvement Program: An Evaluation of All Participating Hospitals. *Annals of Surgery*, *250*(3), 363–376. https://doi.org/10.1097/SLA.0b013e3181b4148f

9. Ingraham, A. M., Richards, K. E., Hall, B. L., & Ko, C. Y. (2010). Quality Improvement in Surgery: The American College of Surgeons National Surgical Quality Improvement Program Approach. *Advances in Surgery*, *44*(1), 251–267. https://doi.org/10.1016/j.yasu.2010.05.003

10. Khuri, S. F., Henderson, W. G., Daley, J., Jonasson, O., Jones, R. S., Campbell, D. A., Fink, A. S., Mentzer, R. M., & Steeger, J. E. (2007). The Patient Safety in Surgery Study: Background, Study Design, and Patient Populations. *Journal of the American College of Surgeons*, *204*(6), 1089–1102. https://doi.org/10.1016/j.jamcollsurg.2007.03.028

11. Hall, B. L., Richards, K., Ingraham, A., & Ko, C. Y. (2009). New approaches to the National Surgical Quality Improvement Program: The American College of Surgeons experience. *The American Journal of Surgery*, *198*(5), S56–S62. https://doi.org/10.1016/j.amjsurg.2009.07.026

12. Shiloach, M., Frencher, S. K., Steeger, J. E., Rowell, K. S., Bartzokis, K., Tomeh, M. G., Richards, K. E., Ko, C. Y., & Hall, B. L. (2010). Toward Robust Information: Data Quality and Inter-Rater Reliability in the American College of Surgeons National Surgical Quality Improvement Program. *Journal of the American College of Surgeons*, *210*(1), 6–16. https://doi.org/10.1016/j.jamcollsurg.2009.09.031

13. Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the Class Imbalance Problem. *2008 Fourth International Conference on Natural Computation*, *4*, 192–201. https://doi.org/10.1109/ICNC.2008.871

14. Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)*, 111–117.

15. Seger, C. (2018). *An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing*. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-237426

16. Cohen, M. E., Ko, C. Y., Bilimoria, K. Y., Zhou, L., Huffman, K., Wang, X., Liu, Y., Kraemer, K., Meng, X., Merkow, R., Chow, W., Matel, B., Richards, K., Hart, A. J., Dimick, J. B., & Hall, B. L. (2013). Optimizing ACS NSQIP Modeling for Evaluation of Surgical Quality and Risk: Patient Risk Adjustment, Procedure Mix Adjustment, Shrinkage Adjustment, and Surgical Focus. *Journal of the American College of Surgeons*, *217*(2), 336-346.e1. https://doi.org/10.1016/j.jamcollsurg.2013.02.027

17. Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

18. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

19. Krzywinski, M., & Altman, N. (2014). Nonparametric tests. *Nature Methods*, *11*(5), 467–468. https://doi.org/10.1038/nmeth.2937

20. Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

21. Ojala, M., & Garriga, G. C. (2009). Permutation Tests for Studying Classifier Performance. *2009 Ninth IEEE International Conference on Data Mining*, 908–913. https://doi.org/10.1109/ICDM.2009.108