Title

Ensemble machine learning for the prediction of patient-level outcomes following thyroidectomy.

Journal

American Journal of Surgery, 222(2)

Authors

Seib, Carolyn

Roose, James

Hubbard, Alan

et al.

Publication Date

2021-08-01

DOI

10.1016/j.amjsurg.2020.11.055

Peer reviewed

HHS Public Access

Author manuscript

Am J Surg. Author manuscript; available in PMC 2022 August 01.

Published in final edited form as:

Am J Surg. 2021 August ; 222(2): 347–353. doi:10.1016/j.amjsurg.2020.11.055.

Ensemble Machine Learning for the Prediction of Patient-Level Outcomes Following Thyroidectomy

Carolyn D. Seib, MD, MAS[1,2], **James P. Roose, MA**[3], **Alan E Hubbard, PhD**[3], **Insoo Suh, MD**[4]

[1.]Stanford–Surgery Policy Improvement Research and Education Center (S-SPIRE), Department of Surgery, Stanford University School of Medicine, Stanford, CA.

[2.]Division of General Surgery, Palo Alto Veterans Affairs Health Care System

[3.]University of California, Berkeley, Division of Biostatistics

[4.]University of California, San Francisco, Section of Endocrine Surgery

Abstract

Background: Accurate prediction of thyroidectomy complications is necessary to inform treatment decisions. Ensemble machine learning provides one approach to improve prediction.

Methods: We applied the Super Learner (SL) algorithm to the 2016–2018 thyroidectomy-specific NSQIP database to predict complications following thyroidectomy. Cross-validation was used to assess model discrimination and precision.

Results: For the 17,987 patients undergoing thyroidectomy, rates of recurrent laryngeal nerve injury, post-operative hypocalcemia prior to discharge or within 30 days, and neck hematoma were 6.1%, 6.4%, 9.0%, and 1.8%, respectively. SL improved prediction of thyroidectomy-specific outcomes when compared with benchmark logistic regression approaches. For postoperative hypocalcemia prior to discharge, SL improved the cross-validated AUROC to 0.72 (95%CI 0.70–0.74) compared to 0.70 (95%CI 0.68–0.72; $p<0.001$) when using a manually curated logistic regression algorithm.

Conclusion: Ensemble machine learning modestly improves prediction for thyroidectomy-specific outcomes. SL holds promise to provide more accurate patient-level risk prediction to inform treatment decisions.

Keywords

Thyroidectomy; machine learning; surgical risk prediction

Corresponding Author: Carolyn Dacey Seib, MD, MAS, Stanford University, 300 Pasteur Drive, H3680, Stanford, CA 94305, Mobile: 917-747-4782, cseib@stanford.edu.

Conflict of Interest Disclosures: There are no conflicts of interest specific to this study. Dr. Suh serves as a consultant for Medtronic and Prescient Surgical. Mr. Roose is an employee of Flatiron Health, Inc.

INTRODUCTION

Estimating the risk of complications following thyroidectomy is an area of great interest that has the potential to inform patient decision-making and improve patient outcomes. The proliferation of electronic health records and large, multidimensional patient databases has provided medical professionals with unprecedented amounts of clinical data to analyze and has created challenges for commonly used statistical methods. Traditional outcomes research using patient-level data has focused on the relationship between preoperative characteristics and postoperative outcomes using parametric statistical methods, such as logistic or linear regression.[1–4] In addition, clinical tools, such as the American College of Surgeons (ACS) patient risk calculator, often use these methodologies to predict the risk of 30-day morbidity and mortality for the purpose of decision-making related to a wide array of operations, with mixed predictive ability.[5,6] A relatively small number of publications have begun to incorporate machine learning technology in their analytic arsenal to estimate the risk of perioperative complications.[7,8]

Super Learner is a cross validation-based, ensemble machine learning approach that has been used to predict patient outcomes following traumatic injury and neo-natal operations, in addition to responses to Human Immunodeficiency Virus therapy,[91011] but has not been applied to risk prediction in endocrine surgery. Traditional parametric regression techniques have limitations in predictive performance, managing non-linear relationships for a large number of predictors, and handling interaction effects. Machine learning approaches are more flexible and can capture more complex functional relationships between predictors and outcome. This can improve predictive performance compared with parametric regression models, especially when the assumptions of the parametric regression model are not met.[12] Ensemble machine learning methods incorporate multiple learning algorithms into modeling to improve predictive performance. Super Learner, as an ensemble machine learning framework, uses an extensive array of standard, sophisticated and flexible machine learning algorithms, including tree-based data adaptive methods, to analyze all available clinical data and, via cross-validation, identify the optimal prediction algorithm among all weighted combinations of the candidate algorithms.[12] The goal of this methodology is to minimize prediction error due to model misspecification by using the data itself to determine the relationship between the predictors and the outcome, without imposing strict assumptions about that relationship.[12]

Risk prediction in thyroid surgery is important to inform patient and provider treatment decisions and to guide perioperative patient management. The most common endocrine-specific complications of thyroid surgery include postoperative neck hematoma, recurrent laryngeal nerve injury, and symptomatic hypocalcemia. While these complications are generally rare, patient factors that contribute to increased risk may result in the risks of elective surgery outweighing its benefits, or inform perioperative interventions to prevent complications, such as the administration of prophylactic calcium or calcitriol. Previous studies focused on thyroidectomy in the ACS National Surgical Quality Improvement Program (NSQIP) database have used traditional statistical methods like logistic regression to estimate associations between pre-operative patient characteristics and composite outcome measures.[1,13,14] The applicability of 30-day outcomes documented in the general

ACS NSQIP participant use file (PUF) to patients undergoing lower risk operations, such as thyroidectomy, has been questioned.[15] As a result, in 2016 the NSQIP PUF was supplemented with additional pre-operative measurements and complication data specific to thyroidectomy that were previously unavailable. Few studies have been published using the thyroidectomy-specific NSQIP participant use file and none of these have focused on patient-specific risk prediction or utilized ensemble machine learning.

In this study, we aimed to determine the accuracy of risk prediction for patients undergoing thyroidectomy using the ensemble machine learning platform Super Learner applied to the NSQIP thyroidectomy-specific PUF. Our goal was to use Super Learner to identify patients at high risk of neck hematoma, recurrent laryngeal nerve injury, and hypocalcemia prior to hospital discharge or within 30 days following thyroidectomy. In doing this, we sought to assess the overall predictive capabilities of data within the NSQIP thyroidectomy-specific PUF for endocrine-specific complications. We hypothesized that Super Learner would produce improved patient-level risk prediction compared to benchmark logistic regression models based on Receiver Operating Characteristic (ROC) curves, precision Recall (PR) curves, and positive predictive value (PPV) in identifying high-risk patients.

METHODS

Data and Patient Population

Patient data used for our analysis included the ACS NSQIP general and procedure-targeted thyroidectomy PUFs from 2016 to 2018. The ACS NSQIP PUF provides validated patient data on a large number of preoperative patient characteristics and 30-day perioperative outcomes from participating medical centers in the U.S. and has been described in detail previously.[16,17] The two data sources were merged by the unique CASEID identifier, and only the subset of the general PUFs patients that were present in the thyroidectomy PUFs were used in the analysis. Patients undergoing thyroidectomy were identified by the following Current Procedural Terminology (CPT) codes: 60210, 60212, 60220, 60225, 60240, 60252, 60254, 60260, 60270, and 60271 (eTable 1). This study was deemed exempt from approval by the University of California, San Francisco Institutional Review Board, because it involved analysis of deidentified patient data.

Pre-operative Variables and Outcomes, Data Processing

A total of 79 pre-operative variables are used in prediction. These include patient and disease characteristics likely to influence the occurrence of postoperative complications (i.e. age, indication for surgery, presence of hyperthyroidism, thyroid malignancy). Missing values for preoperative values were imputed using the median for continuous variables, and by sampling from the empirical distribution of available values for categorical variables. Corresponding indicator variables were added for any variable that was imputed to distinguish imputed values from true observed values in prediction. After including the missingness indicators, a total of 120 variables are used for prediction. Four post-operative outcome variables were considered: neck hematoma, recurrent laryngeal nerve injury, hypocalcemia prior to discharge, and hypocalcemia within 30 days of discharge. Patients who underwent thyroid lobectomy (CPT codes 60210, 60220) were excluded from summary

statistics and prediction models related to hypocalcemia outcomes. The risk of recurrent laryngeal nerve injury was considered on a per-patient (rather than per-nerve) basis. Patients for whom an outcome was missing were excluded from the analysis when that outcome was of interest.

Super Learner and Prediction Algorithms

Primary analyses utilized Super Learner, an ensemble machine learning algorithm for selecting the optimal regression or classification algorithm from a set of weighted combinations of proposed candidate algorithms ("the library") using V-fold cross-validation.[18] Super Learner takes as input a matrix of predictor variables, a vector of outcomes, a user-specified library of prediction algorithms (called "learners"), and a loss function, which determines the relative importance of different prediction errors. The model training aims to minimize this loss function. Super Learner uses nested V-fold cross-validation to identify and evaluate the optimal combination of learners (Figure 1). In the "outer" cross-validation, V −1 folds of the data (i.e. 90% of observations for V = 10) are used to train the different learners and choose the optimal way to combine the outputs from the learners, and the remaining 1 fold is used to assess prediction performance. This is repeated such that each fold is withheld from the training process and used to assess performance once. Within the V − 1 "training" folds is the "nested" or "inner" cross-validation process, in which the V-1 training folds are further subdivided into U folds. Then U − 1 folds of the data are used to train the individual learners (for V = 10 and U = 20, this means that 85.5% of the observations are used to train the candidate learners), and the remaining inner fold is used to determine the combination of component learners that minimizes average loss (also known as "cross-validated risk"). This inner cross-validation process is repeated such that each inner validation fold is used to determine the optimal combination of learners once, and the outer cross-validation process is repeated such that each outer fold is used as the validation fold to evaluate performance.[19] The statistical theory related to loss function-based cross-validation indicates that Super Learner will, asymptotically, perform as well as the optimal learner (with respect to the loss function specified) among all possible combinations of candidate learners.[19] Consequently, provided the set of candidate learners is sufficiently large and varied, Super Learner's prediction performance will approach the upper limit of performance for a given prediction problem. This provides a useful benchmark for simpler and less computationally-intensive learners, which in some settings may be more practical or interpretable.

In this study, Super Learner was used to output a predicted probability of the specified complication for each patient and model discrimination was assessed using ROC and PR curves. The loss function used was the negative of the Area Under the ROC Curve (AUROC), meaning that the average of the AUROC was maximized when identifying the optimal combination of candidate learners. The following algorithms were included as learners in the Super Learner library: logistic regression, logistic regression with forward step-wise variable selection,[20] least absolute shrinkage and selection operator (LASSO) logistic regression,[21] elastic-net regularized logistic regression,[22] ridge logistic regression,[23] recursive-partitioning tree,[20] a pruned recursive-partitioning tree,[20] gradient-boosting machine using the xgBoost implementation,[24] conditional inference tree-based random

forest,[25] and a recursive-partitioning based random forest.[26] The aforementioned algorithms are included both with and without the use of several pre-screening approaches that filter the variables used for prediction. The following pre-screening approaches are used: a random forest variable importance filter, a filter to remove near-constant variables, and a custom filter to select only the pre-treatment variables used in a recent study[14] plus the thyroid-specific input variables. The outer V-fold cross-validation process uses V = 10 folds for performance evaluation, and the inner nested cross-validation process uses U = 20 folds for training and construction of the Super Learner. To ensure sufficient number of outcomes in the validation sample, the cross-validation process was stratified by the outcome.

Comparison Logistic Regression Algorithms

Previous work has benchmarked Super Learner's prediction performance against existing risk scoring systems. Because no standard risk scoring system is available for thyroidectomy and our outcomes of interest, two main-term logistic regression models were used as benchmarks. The first benchmark generalized linear model (the "Full GLM") is a logistic regression model that uses *all* the pre-operative variables except those that exhibit insufficient variability to allow cross- validation to be used for model fitting and performance evaluation (meaning that they are nearly constant). The second benchmark generalized linear model (the "Curated GLM") is a logistic regression model that includes only the pre-treatment variables used in a recent study of overall-postoperative complications[14] plus the pre-treatment thyroidectomy-specific variables made available by NSQIP beginning in 2016. AUROC and precision recall curves (AUPRC) were compared to those of benchmark logistic regression methods to assess model discrimination and precision, the latter of which is important to assess model performance in predicting rare outcomes. Confidence intervals and p-values were obtained for AUROC estimates and comparisons of AUROC using an influence function-based methodology and the Delta method.[27] Confidence intervals for AUPRC estimates were obtained using a logit transformation methodology.[28] PPV, negative predictive value (NPV), sensitivity and specificity were used to evaluate classifier performance at a specific threshold. Given the outcomes of interest are rare, a threshold of 15% was chosen to assess performance of identifying "high-risk" patients. Equality in PPV at the 15% threshold between Super Learner and each benchmark logistic regression was tested using the weighted generalized score statistic.[29]

RESULTS

Population Characteristics and Thyroidectomy Complications

We identified a total of 17,987 patients who underwent thyroidectomy in the 2016–2018 ACS NSQIP thyroidectomy-specific PUFs. In this population, 14,004 (77.9%) were female and 64% of patients underwent outpatient thyroid surgery. A total of 8,205 (45.6%) underwent total or subtotal thyroidectomy, 6,605 (36.7%) thyroid lobectomy, 2,268 (12.6%) thyroidectomy with limited or radical neck dissection, and 909 (5.1%) reoperative thyroid surgery. Complete baseline demographic and preoperative characteristics of the study population are listed in Table 1. The incidence of neck hematoma, recurrent laryngeal nerve injury, and hypocalcemia prior to discharge and at 30 days are listed in Table 2.

Performance of Super Learner: AUROC and AUPRC

Super Learner improved prediction performance based on both the AUROC and the area under the PR Curve (AUPRC) for all outcomes based on comparisons of point estimates (Table 3). The best prediction performance and greatest improvement against the benchmarks was achieved for the post-operative hypocalcemia prior to discharge, for which Super Learner achieved an AUROC of 0.720 (95% CI: 0.702 – 0.739) compared with the Full GLM (0.711, 95% CI: 0.692 – 0.731; p=0.027) and Curated GLM (0.704, 95% CI: 0.684 – 0.723; p<0.001). For the other outcomes, the absolute performance and performance relative to the benchmark were similarly modest. ROC curves for all four outcomes and all three prediction approaches are included as Figure 2.

The AUPRC summarizes PPV across the range of possible sensitivity levels and has been promoted as a more appropriate measure of prediction when considering rare outcomes. By focusing on PPV and sensitivity, AUPRC reveals how effectively a classifier can identify the rare positive outcomes. This avoids the pitfall that occurs when a high AUROC can be driven by improved performance in the region far to the right of the ROC, where the false positive rate is near 1.[30] Super Learner outperformed the benchmark approaches when measured by the AUPRC, but no method showed good absolute performance by this metric. All three analytic approaches perform better on the two hypocalcemia related outcomes. For hypocalcemia prior to discharge, Super Learner's AUPRC was 0.164 (95% CI: 0.157 – 0.171), compared with 0.161 (95% CI: 0.155 – 0.168) and 0.142 (95% CI: 0.136 – 0.149) for the Full GLM and Curated GLM, respectively. Similarly, for hypocalcemia within 30-days, Super Learner's AUPRC 0.157 (95% CI: 0.150 – 0.164) improved upon that of the two logistic regression approaches. Prediction performance measured by AUPRC is poor for the other outcomes, shown in (Table 3).

Positive Predictive Value and Negative Predictive Value

Using 15% risk of hypocalcemia as our threshold, we calculated the PPV, NPV, sensitivity, and specificity for the three approaches (eTable 2). Super Learner improved the estimated positive predictive value of hypocalcemia before discharge, and at 30 days, to 22% (95% CI 19% – 25%) from 19% (95% CI 16% – 21%; p = 0.004) and to 21% (95% CI 18% – 23%) from 17% (95% CI 14% – 19%;p<0.0005), respectively.[29]

DISCUSSION

In this study, we found that Super Learner improves prediction of post-thyroidectomy complications as measured by AUROC and AUPRC by a modest but statistically significant amount when compared with proposed benchmark logistic regression methods. In absolute terms, the prediction of postoperative hypocalcemia following thyroidectomy was superior to prediction performance for recurrent laryngeal nerve injury and neck hematoma across all three prediction algorithms. For post-operative hypocalcemia prior to discharge, Super Learner increases AUROC by approximately 0.02 compared with the Curated GLM. Using a probabilistic interpretation of these results, an AUROC difference of 0.02 can be interpreted as follows: out of 100 pairs of patients formed of one patient who later develops hypocalcemia prior to discharge, and one who does not, Super Learner would correctly

classify two pairs more than the Curated GLM, on average.[31] These findings suggest ensemble machine learning holds potential for improved prediction of patient-level complications following thyroidectomy but, due to limitations of the NSQIP thyroidectomy-specific PUF, improvements in prediction with these data were not clinically meaningful.

This study is one of the first implementations of machine learning for risk prediction in the field of endocrine surgery. It is also one of the first uses of the thyroidectomy procedure-specific NSQIP data. Given the theoretical basis for Super Learner guarantees that it will perform as well, or better, than the best learner in the Super Learner library, these results can be interpreted as an approximate upper bound for prediction performance in thyroidectomy using the data available in the thyroidectomy procedure-specific NSQIP PUF. Although additional learners, different pre-processing and imputation approaches, or additional data could all improve the performance of Super Learner, the fact that the absolute performance measured by AUROC is poor or modest for all four outcomes suggests the ACS NSQIP thyroidectomy-specific PUF does not provide enough patient-specific data to reliably predict the rare outcomes of neck hematoma, recurrent laryngeal nerve injury, and postoperative hypocalcemia. Improved identification of patients who are high-risk for neck hematoma or hypocalcemia may improve clinician decision-making about closer patient monitoring to prevent adverse outcomes and prophylactic administration of calcium or calcitriol, and an accurate assessment of the risk of recurrent laryngeal nerve injury is needed for informed patient decision-making. Further research is needed to determine if the incorporation of more granular clinical information, such as preoperative lab results, intraoperative findings, or even adjunct studies, such as intraoperative nerve monitoring or parathyroid autoflourescence data, can improve risk prediction in thyroid surgery and allow for more informed treatment decisions by patients and providers.

It is generally recognized that prediction with rare outcomes is challenging, and it is difficult to achieve high sensitivity in such settings. Nevertheless, a recent study of neo-natal surgery mortality using Super Learner and NSQIP data achieved an AUROC of 0.91 using a similar library of algorithms, a smaller sample size (6,499 cases), and when focusing on a similarly rare outcome (3.5%)[10]. A previous study of mortality in the ICU also achieved a higher AUROC of 0.85, when focused on a more prevalent outcome (12.2%) and using a larger sample of patients.[9] A review of the uses of machine learning in neurosurgery found that machine learning approaches improved AUROC by, on average, 15% or 0.06, achieving a median AUROC of 0.83.[7] The lower risk of thyroidectomy compared to neo-natal surgery, intensive care unit management of critically ill patients, and neurosurgery, in addition to the size of the NSQIP dataset likely contributes to the relatively poor performance of Super Learner in this study. This suggests that the application of Super Learner to databases with a larger amount of granular clinical data or electronic health record data may be the next step to improve our ability to identify high-risk patients and allow us to counsel them or adjust our treatment plans appropriately.

There is great excitement around the use of big data and machine learning as a tool in clinical decision making, but there are significant challenges to analyzing these data in a clinically meaningful way. In practice, surgeons may rely on only a small number of variables in decision making. However, using software like Super Learner and rigorous,

well-defined variable importance measures that represent statistical parameters of interest offers another approach to incorporating machine learning into clinical practice. Ongoing studies like this one, are needed to familiarize clinicians with the prediction capability of ensemble machine learning algorithms and, eventually, develop easy to use and validated risk prediction software based on this methodology.

The NSQIP database has several well-documented limitations that may affect our prediction abilities. The patient records collected in NSQIP are from a limited number of hospitals (93 in 2016, 91 in 2017, and 112 in 2018), and a small number of surgeons within each hospital, but these are not identifiable within the PUF. As a result, we are unable to account for clustering by institution and provider, which presents an obstacle to valid inference and the generalizability of our results. Variable importance measures could provide insight into the relative association of preoperative variables with the different outcomes,[31] similar to traditional odds ratios, but an optimal approach to understanding variable importance for Super Learner has not yet been established and widely used. In addition, in risk-prediction it is difficult to identify or account for prophylactic vs. therapeutic interventions that may impact outcomes of interest, such as the indications for administration of calcium and vitamin D analogs in the NSQIP procedure-targeted thyroidectomy PUF. In addition to the other shortcomings highlighted above, these variable-related issues should be addressed in future iterations of the thyroidectomy-specific database.

Conclusions and Future Directions

Ensemble machine learning improves prediction of post-thyroidectomy complications compared to traditional parametric multivariate analytic methods using the NSQIP procedure-targeted thyroidectomy PUF, although this improvement was marginal. This highlights limitations of these data and suggests focused efforts are needed to improve the information collected and included in databases meant for quality assessment in thyroid surgery. Super Learner has the potential to provide more robust and accurate patient-level risk prediction to inform treatment decisions related to thyroidectomy, but clinically significant improvements in predictive performance are dependent on the granularity and quality of clinical inputs. As additional technologies are employed to improve the richness of patient-level data collection, further research is needed to learn how to harness this data to improve predictive capabilities in endocrine surgery.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

The American College of Surgeons National Surgical Quality Improvement Program and the hospitals participating in the ACS NSQIP 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.

Funding/Support: This work was supported by the National Institutes of Health, National Institute on Aging [R03AG06009].

REFERENCES

1. Seib CD, Rochefort H, Chomsky-Higgins K, et al. Association of patient frailty with increased morbidity after common ambulatory general surgery operations. JAMA surgery. 2018;153(2):160–168. [PubMed: 29049457]
2. Zhou J, Zhou Y, Cao S, et al. Multivariate logistic regression analysis of postoperative complications and risk model establishment of gastrectomy for gastric cancer: A single-center cohort report. Scandinavian journal of gastroenterology. 2016;51(1):8–15. [PubMed: 26228994]
3. Ozrazgat-Baslanti T, Blanc P, Thottakkara P, et al. Preoperative assessment of the risk for multiple complications after surgery. Surgery. 2016;160(2):463–472. [PubMed: 27238354]
4. Abraham CR, Ata A, Carsello CB, Chan TL, Stain SC, Beyer TD. A NSQIP risk assessment for thyroid surgery based on comorbidities. Journal of the American College of Surgeons. 2014;218(6):1231–1237. [PubMed: 24745620]
5. Bilimoria KY, Liu Y, Paruch JL, et al. 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. 2013;217(5):833–842. [PubMed: 24055383]
6. Vaziri S, Wilson J, Abbatematteo J, et al. Predictive performance of the American College of Surgeons universal risk calculator in neurosurgical patients. Journal of neurosurgery. 2018;128(3):942–947. [PubMed: 28452615]
7. Senders JT, Staples PC, Karhade AV, et al. Machine learning and neurosurgical outcome prediction: a systematic review. World neurosurgery. 2018;109:476–486. [PubMed: 28986230]
8. Ehlers AP, Roy SB, Khor S, et al. Improved Risk Prediction Following Surgery Using Machine Learning Algorithms. eGEMs. 2017;5(2).
9. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. The Lancet Respiratory Medicine. 2015;3(1):42–52. [PubMed: 25466337]
10. Cooper JN, Wei L, Fernandez SA, Minneci PC, Deans KJ. Pre-operative prediction of surgical morbidity in children: comparison of five statistical models. Computers in biology and medicine. 2015;57:54–65. [PubMed: 25528697]
11. Sinisi SE, Polley EC, Petersen ML, Rhee S-Y, Van Der Laan MJ. Super learning: an application to the prediction of HIV-1 drug resistance. Statistical applications in genetics and molecular biology. 2007;6(1).
12. Polley EC, Rose S, van der Laan MJ. Targeted Learning: Casual Inference for Observational and Experimental Data, chapter 3: Super Learning. In: Springer, New York; 2011.
13. Iannuzzi JC, Fleming FJ, Kelly KN, Ruan DT, Monson JR, Moalem J. Risk scoring can predict readmission after endocrine surgery. Surgery. 2014;156(6):1432–1440. [PubMed: 25456927]
14. Caulley L, Johnson-Obaseki S, Luo L, Javidnia H. Risk factors for postoperative complications in total thyroidectomy: A retrospective, risk-adjusted analysis from the National Surgical Quality Improvement Program. Medicine. 2017;96(5).
15. Sippel RS, Chen H. Limitations of the ACS NSQIP in thyroid surgery. Annals of surgical oncology. 2011;18(13):3529–3530. [PubMed: 21755377]
16. Khuri SF, Daley J, Henderson W, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. Annals of surgery. 1998;228(4):491. [PubMed: 9790339]
17. American College of Surgeons National Quality Improvement Program Website. https://www.facs.org/quality-programs/acs-nsqip. Published 2019. Accessed.
18. Polley EC, Van Der Laan MJ. Super learner in prediction. 2010.
19. Van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media; 2011.
20. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer series in statistics. In: :. Springer; 2001.

21. Tibshirani R Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267–288.
22. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301–320.
23. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67.
24. Dombrowsky A, Borg B, Xie R, Kirklin JK, Chen H, Balentine CJ. Why Is Hyperparathyroidism Underdiagnosed and Undertreated in Older Adults? Clinical Medicine Insights: Endocrinology and Diabetes. 2018;11:1179551418815916.
25. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics. 2006;15(3):651–674.
26. Breiman L Random forests. Machine learning. 2001;45(1):5–32.
27. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Electronic journal of statistics. 2015;9(1):1583. [PubMed: 26279737]
28. Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. 2013.
29. Kosinski AS. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. Statistics in medicine. 2013;32(6):964–977. [PubMed: 22912343]
30. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one. 2015;10(3):e0118432. [PubMed: 25738806]
31. Díaz I, Hubbard A, Decker A, Cohen M. Variable importance and prediction methods for longitudinal problems with missing variables. PloS one. 2015;10(3):e0120031. [PubMed: 25815719]

HIGHLIGHTS

- Postoperative hypocalcemia within 30 days is the most common complication following thyroidectomy.
- Ensemble machine learning modestly improved prediction of thyroidectomy-specific outcomes.
- Use of machine learning in databases with granular clinical information holds promise to improve patient-level risk-prediction and counseling.

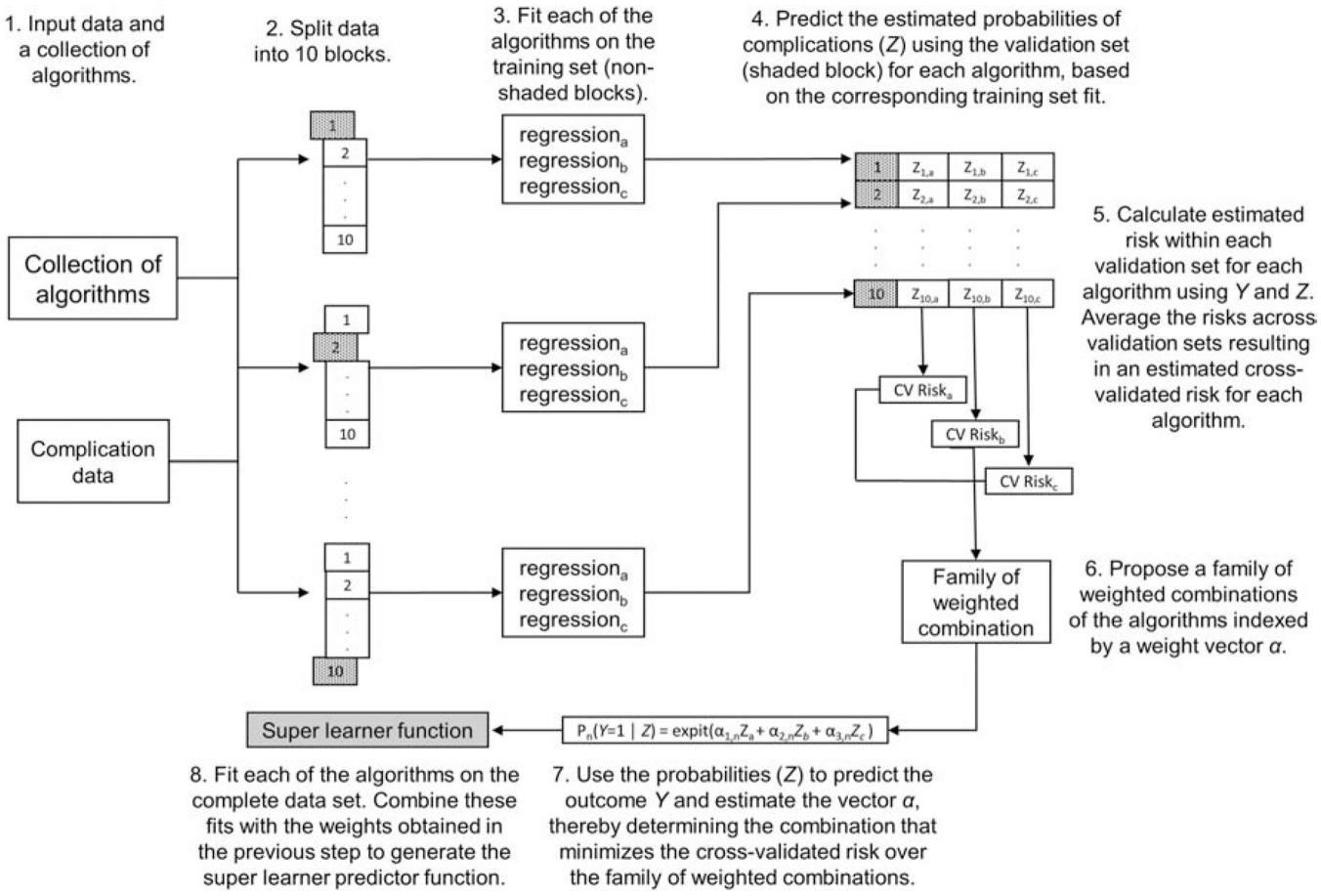

Figure 1:

Illustration of the inner cross validation of Super Learner for predicting thyroidectomy complications. Adapted with permission from Springer Nature from *Targeted Learning*: Super Learning. Eric C. Polley; Sherri Rose; Mark J. van der Laan, 2011.[19]

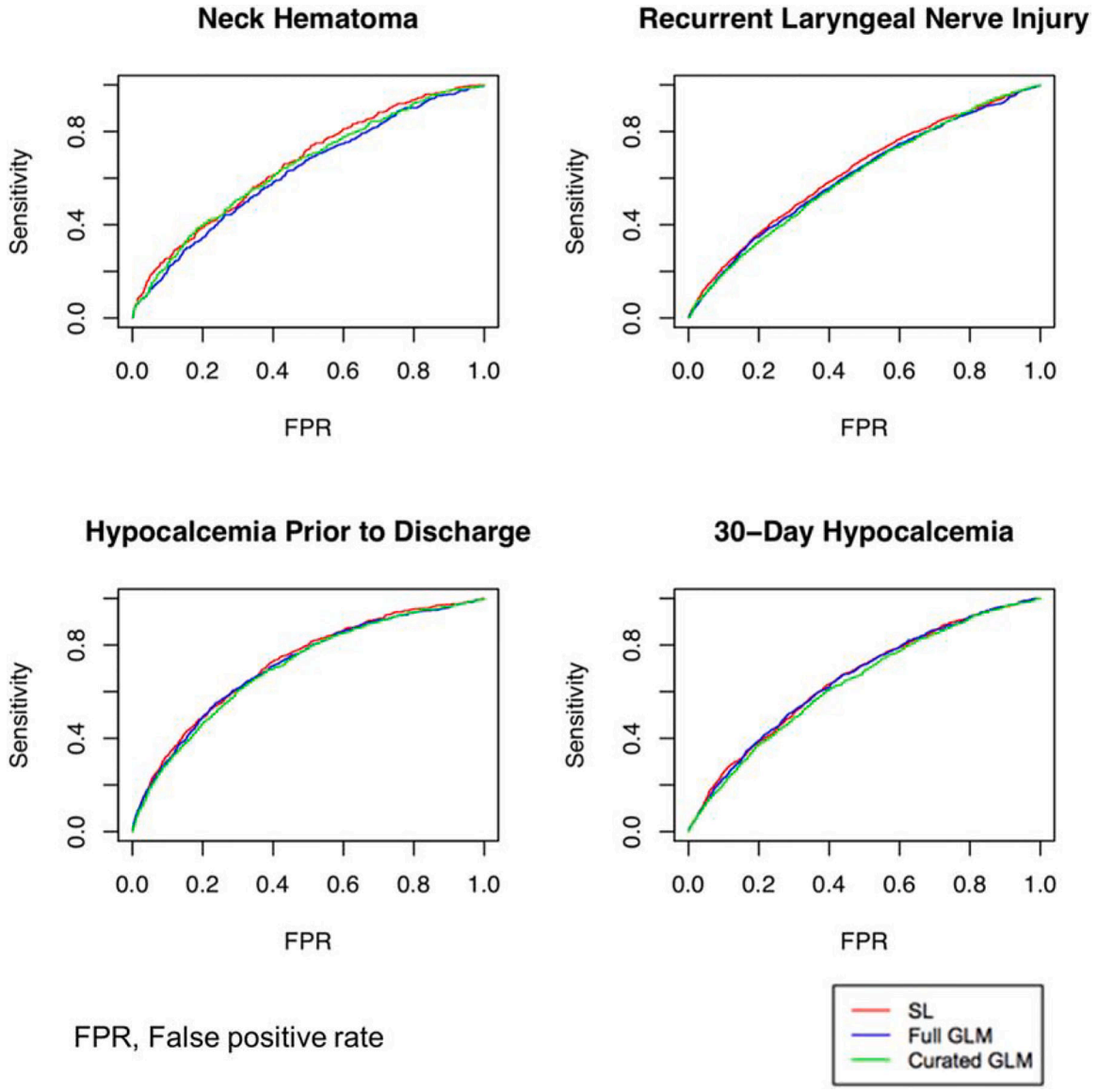

Figure 2:

Receiver Operator Characteristic Curves for Super Learner, the Full Generalized Linear Model (GLM) and Curated GLM for thyroidectomy-specific outcomes.

Table 1:

Baseline characteristics of patients undergoing thyroid surgery in the 2016 to 2018 NSQIP Procedure-Targeted Thyroidectomy PUF.

Characteristic	All patients (n = 17,987)
Demographics	
Age—mean (SD)	51.7 (15.0)
Female gender—no. (%)	14,004 (77.9)
Race—no. (%)	
White	9,908 (62.8)
Asian or Pacific Islander	1,012 (6.4)
Black or African American	1,786 (11.3)
American Indian or Alaskan Native	75 (0.5)
Other or Unknown	2,989 (19.0)
Preoperative Health & Comorbidities	
Body mass index, kg/m2—mean (SD)	30.6 (7.6)
Weight loss (>10% in 6 months)—no. (%)	111 (0.6)
Current smoker—no. (%)	2,581 (14.3)
Diabetes mellitus—no. (%)	2,415 (13.4)
Chronic obstructive pulmonary disease—no. (%)	454 (2.5)
Hypertension requiring medication—no. (%)	6,879 (38.2)
Steroid use—no. (%)	506 (2.8)
Congestive heart failure—no. (%)	76 (0.4)
Functional status prior to surgery— no. (%)	
Independent	17,888 (99.4)
Partially or totally dependent	91 (0.5)
Unknown	8 (0.0)
ASA Class—no. (%)	
No or mild systemic disease	11,460 (63.7)
Severe systemic disease	6,052 (33.6)
Life threatening systemic disease or moribund	376 (2.1)237 (2.0)
None Assigned	99 (0.6)
Type of Operation	
Total or subtotal thyroidectomy—no. (%)	8,205 (45.6)
Thyroid lobectomy—no. (%)	6,605 (36.7)
Thyroidectomy with limited or radical neck dissection—no. (%)	2,268 (12.6)
Reoperative thyroid surgery—no. (%)	909 (5.1)

Table 2:

Summary of Thyroidectomy-Specific Outcomes of Interest.

Outcome	Total number of operations*	Operations with complications	Complication Rate
Neck hematoma	17,831	325	1.8%
Recurrent laryngeal nerve injury	17,784	1,087	6.1%
Hypocalcemia prior to discharge	11,259	723	6.4%
Hypocalcemia within 30 days	10,995	985	9.0%

*The number of operations considered for each outcome differed because operations missing each thyroidectomy-specific outcomes were excluded for that calculation, and patients who underwent thyroid lobectomy (CPT codes 60210, 60220) were excluded from summary statistics and prediction models related to hypocalcemia outcomes.

Table 3:

Summary of AUROC and AUPRC Results for All Three Algorithms.

Outcome	Super Learner	Full GLM	Curated GLM
AUROC (95% CI)			
Neck hematoma	0.663 (0.634 – 0.691)	0.627 (0.597 – 0.657)	0.648 (0.619 – 0.678)
Recurrent laryngeal nerve injury	0.628 (0.611 – 0.646)	0.612 (0.594 – 0.629)	0.608 (0.591 – 0.625)
Hypocalcemia prior to discharge	0.720 (0.702 – 0.739)	0.711 (0.692 – 0.731)	0.704 (0.684 – 0.723)
Hypocalcemia within 30 days	0.656 (0.638 – 0.673)	0.655 (0.637 – 0.672)	0.638 (0.620 – 0.655)
AUPRC (95% CI)			
Neck hematoma	0.040 (0.037 – 0.043)	0.032 (0.030 – 0.035)	0.036 (0.033 – 0.039)
Recurrent laryngeal nerve injury	0.103 (0.099 – 0.108)	0.094 (0.090 – 0.098)	0.090 (0.086 – 0.095)
Hypocalcemia prior to discharge	0.164 (0.157 – 0.171)	0.161 (0.155 – 0.168)	0.142 (0.136 – 0.149)
Hypocalcemia within 30 days	0.157 (0.150 – 0.164)	0.154 (0.147 – 0.161)	0.143 (0.136 – 0.149)