A machine learning-based approach for individualized prediction of short-term outcomes after anterior cervical corpectomy

Mert Karabacak, Alexander Schupper, Matthew Carr, Konstantinos Margetis

Department of Neurosurgery, Mount Sinai Health System, New York, NY, USA

Received Feb 5, 2024; Revised Mar 20, 2024; Accepted Apr 15, 2024

Corresponding Author: Konstantinos Margetis
Department of Neurosurgery, Mount Sinai Health System, New York, NY, USA
Tel: +1-212-241-3649, Fax: +1-212-410-0603, E-mail: konstantinos.margetis@mountsinai.org

Study Design: A retrospective machine learning (ML) classification study for prognostic modeling after anterior cervical corpectomy (ACC).

Purpose: To evaluate the effectiveness of ML in predicting ACC outcomes and develop an accessible, user-friendly tool for this purpose.

Overview of Literature: Based on our literature review, no study has examined the capability of ML algorithms to predict major short-term ACC outcomes, such as prolonged length of hospital stay (LOS), non-home discharge, and major complications.

Methods: The American College of Surgeons' National Surgical Quality Improvement Program database was used to identify patients who underwent ACC. Prolonged LOS, non-home discharges, and major complications were assessed as the outcomes of interest. ML models were developed with the TabPFN algorithm and integrated into an open-access website to predict these outcomes.

Results: The models for predicting prolonged LOS, non-home discharges, and major complications demonstrated mean areas under the receiver operating characteristic curve (AUROC) of 0.802, 0.816, and 0.702, respectively. These findings highlight the discriminatory capacities of the models: fair (AUROC >0.7) for differentiating patients with major complications from those without, and good (AUROC >0.8) for distinguishing between those with and without prolonged LOS and non-home discharges. According to the SHapley Additive Explanations analysis, single- versus multiple-level surgery, age, body mass index, preoperative hematocrit, and American Society of Anesthesiologists physical status repetitively emerged as the most important variables for each outcome.

Conclusions: This study has considerably enhanced the prediction of postoperative results after ACC surgery by implementing advanced ML techniques. A major contribution is the creation of an accessible web application, highlighting the practical value of the developed models. Our findings imply that ML can serve as an invaluable supplementary tool to stratify patient risk for this procedure and can predict diverse postoperative adverse outcomes.

Keywords: Cervical vertebrae; Degenerative spine disease; Anterior cervical corpectomy; Machine learning

Introduction

Anterior cervical corpectomy (ACC) with fusion is performed to treat various cervical spine disorders, including degenerative diseases, traumas, neoplasms, and infections. Aside from ACC with fusion, anterior cervical spine surgery approaches also include anterior cervical discectomy and fusion (ACDF), and both
have been associated with favorable clinical outcomes [1,2]. Compared with ACC, ACDF can better preserve the biomechanical stability of the cervical spine while decreasing the likelihood of complications, such as implant displacement [3]. Moreover, potential limitations of ACDF comprise incomplete decompression, spinal cord injury, limited surgical exposure, and increased incidence of pseudarthrosis [4,5]. Conversely, ACC is more prone to dural tears, cerebrospinal fluid leakage, bone graft displacement, and other complications [6,7]. In degenerative diseases, corpectomy is indicated for compression behind the vertebral body beyond the disc space level. It permits adequate spinal cord decompression in such cases. ACC allows for direct and adequate decompression of the anterior spinal cord, resection of lesions involving the cervical vertebral bodies, and enhancement of the cervical lordosis [8]. Given the potential benefits and drawbacks of this technique, preoperatively determining patients likely to have post-ACC adverse events becomes critical.

The monitoring and generation of risk-adjusted estimates for adverse postoperative outcomes aim to curb healthcare costs. Consequently, clinicians must process increasing volumes of complex data, necessitating more advanced analytical techniques [9]. Machine learning (ML)-based predictive models can effectively leverage high-dimensional clinical data to create precise patient risk assessment tools. ML has been successfully utilized to predict short-term outcomes after various spinal conditions and procedures [10-13]. However, based on our literature review, no study has examined the capability of ML algorithms to predict major short-term adverse events after ACC, such as prolonged length of hospital stay (LOS), non-home discharges, and major complications. Thus, this study aimed to evaluate the effectiveness of ML in predicting ACC outcomes and develop an accessible, user-friendly tool for this purpose.

Materials and Methods

Ethical approval

This study was conducted in compliance with the principles of the Declaration of Helsinki. This study was deemed exempt from review by the institutional review board of Icahn School of Medicine at Mount Sinai (STUDY-22-01302-MOD002) because it involved analysis of deidentified patient data. Informed consent was not required given the retrospective analysis of previously collected deidentified information.

Study design and guidelines

This retrospective ML classification study (with binary categorical outcomes) for prognostic modeling after ACC followed the Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis [14] and Journal of Medical Internet Research Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [15] guidelines.

Data source

The American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) database was examined to identify patients undergoing ACC between 2014 and 2020. ACS-NSQIP is a national registry comprising data on major surgeries across all specialties performed on adult patients at >700 participating medical centers in the United States [16,17]. The structure and data collection methods of the database have been discussed extensively elsewhere [18].

Study population

The NSQIP database was queried to identify patients who met the following inclusion criteria: (1) Common Procedural Terminology (CPT) code 63081: vertebral corpectomy (vertebral body resection), partial or complete, anterior approach with decompression of the spinal cord and/or nerve root(s), (2) elective surgery, (3) surgery under general anesthesia, and (4) surgical subspecialty of neurosurgery or orthopedics. Patients were excluded based on the following criteria: (1) emergency surgery, (2) unclean wounds (wound classes 2–4), (3) patients with sepsis/shock/systemic inflammatory response syndrome 48 hours before surgery, (4) American Society of Anesthesiologists (ASA) physical status classification score ≥4 or unassigned, (5) LOS exceeding the 30-day postoperative period captured in the database, and (6) patients who died in hospital, left against medical advice, or discharged to hospice. Those undergoing concurrent cervical surgeries, thoracic and lumbar fusions, revision procedures, or intraspinal lesion resections were also excluded. Supplement 1 provides the CPT codes for these exclusionary surgeries.

Predictor variables

The predictor variables included in the analysis consisted of data that would have been known before the out-
comes of interest following surgery. The predictors fell into four categories: (1) demographics, including age, sex, race, Hispanic ethnicity, body mass index (BMI), and transfer status; (2) comorbidities, namely, diabetes mellitus, smoking within a year, dyspnea, severe chronic obstructive pulmonary disease history, congestive heart failure, hypertension, acute kidney injury, currently requiring or on dialysis, disseminated cancer, steroids or immunosuppressants for a chronic disease, >10% weight loss over 6 months, bleeding disorders, ≥1 unit of red blood cell transfusion 72 hours before surgery, ASA physical status classification, and preoperative functional status; (3) preoperative laboratory values, including hematocrit, white blood cell count, platelet count, prothrombin time, international normalized ratio (INR), partial thromboplastin time (PTT), serum sodium, blood urea nitrogen, serum creatinine, serum albumin, total bilirubin, serum glutamic-oxaloacetic transaminase (SGOT), and serum alkaline phosphatase (ALP); and (4) operative details, such as surgical specialty, single- versus multiple-level surgery (CPT code 63082). Definitions for predictor variables are provided in the ACS-NSQIP participant user data file guides (www.facs.org/quality-programs/data-and-registries/acs-nsqip/participant-use-data-file).

Three outcomes were analyzed, namely, prolonged LOS, non-home discharge, and occurrence of any major complications. Prolonged LOS was defined as LOS exceeding 80% of the total patient cohort, corresponding to ≥4 days. Discharge destinations were classified as home versus non-home. Non-home discharges included facilities requiring post-hospital care such as rehabilitation, skilled nursing, separate acute inpatient stay, unskilled care, or senior living communities. Home discharges included those discharged to a facility serving as the residence or actual home of the patients. Major postoperative complications comprised one or more of the following: deep surgical site infections, wound dehiscence, need for reintubation, pulmonary embolism, prolonged mechanical ventilation >48 hours, renal insufficiency or failure requiring dialysis, cardiac arrest, myocardial infarction, bleeding requiring ≥1 unit of red blood cell transfusion, deep venous thrombosis, sepsis, and septic shock. The database also recorded less severe complications such as superficial surgical site infections, pneumonia, and urinary tract infections. However, these were not classified as major for this analysis. Any patients with missing data for prolonged LOS, non-home discharge, or major complications were excluded when analyzing that outcome.

Data preprocessing and partition

Imputation methods were used to substitute missing values and prevent potential bias from excluding patients with incomplete data. For continuous variables, missing values were imputed using the k-nearest neighbor algorithm when <25% of data were missing [19]. Any continuous variables with >25% missing values were excluded (prothrombin time, INR, PTT, serum albumin, total bilirubin, SGOT, and serum ALP). For categorical variables, missing values were filled as “unknown” or “unknown/other”; if again <25% of data were missing, the variable was excluded from the analysis (smoking within a year).

To address potential class imbalance in the training data, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized before model training. SMOTE resolves skewed class distributions by synthetically generating new cases belonging to the minority class rather than duplicating available samples. This technique expands the number of cases from under-represented groups and improves model performance compared with simply replicating minority observations [20]. Applying SMOTE ensured adequate instances of all classes and prevented learning bias during training that tends to favor majority groups.

Model development and evaluation

TabPFN, a modified prior-data fitted network architecture, was used to develop the prediction models [21,22]. Prior-data fitted networks, including TabPFN, are pre-trained on synthetic data to emulate Bayesian inference on real-world information. TabPFN uses a meta-learning approach to adapt to new, unseen data by training across different datasets [22,23]. This pre-training enables TabPFN to learn complex patterns in tabular data and smoothly transition when applied to novel datasets.

A fivefold stratified cross-validation approach was used to evaluate model performance. The data were divided into five roughly equal partitions, balancing the ratio of outcome classes across folds (stratification) to ensure consistent class representation. In each fold, the initial training set (80% of total data) was further split into final training (70% of total data) and validation (10% of total data) subsets. This step produced a 70:10:20 ratio for training to validation to holdout testing. The validation subsets enabled isotonic calibration to match predicted risks with actual outcomes [24]. scikit-learn’s “CalibratedClassifierCV” class handled this sigmoid calibration fitting on the validation data [25].

The calibrated TabPFN model generated predictions
and probability estimates on each test fold. Overall performance was evaluated by aggregating the results across all folds because cross-validation allows reliable assessment of generalizable predictive ability. For enhanced interpretability, SHapley Additive Explanation (SHAP) values were used to determine the relative importance of features. The SHAP plot displayed the variables chosen hierarchically, with the most impactful at the top. To enable full transparency, the model code is publicly available in the study GitHub repository (https://github.com/mertkarabacak/NSQIP-ACC).

Model performance was assessed visually using a receiver operating characteristic (ROC) curve showing the tradeoff between true- and false-positive rates across thresholds and a precision–recall curve (PRC) illustrating the precision–recall balance. Several metrics were calculated: weighted precision (ratio of true positives to total positive predictions weighted by class frequencies), weighted recall (ratio of true positives to total actual positives weighted class frequencies), F1 score (harmonic mean of precision and recall), area under the ROC curve (AUROC, evaluates class discrimination across thresholds), area under the PRC (AUPRC, reflects precision–recall tradeoff across thresholds), and Brier score (mean squared difference between predicted probabilities and actual outcomes). A 95% bootstrap confidence interval with 1,000 resampled datasets was calculated for each metric. This involved sampling with replacement to generate 1,000 new samples, with the 2.5th and 97.5th percentiles of the bootstrap distribution to determine the 95% confidence intervals.

Web application

A web application was developed to facilitate patient-level outcome predictions for each outcome. The application integrated these trained models with the implementation source code shared publicly on Hugging Face, which is a platform enabling model distribution. The application can be accessed at https://huggingface.co/spaces/MSHS-Neurosurgery-Research/NSQIP-ACC.

Descriptive statistics

For continuous variables following a normal distribution, means±standard deviations were reported, whereas medians (interquartile ranges) were presented for non-normally distributed variables. Categorical variables are presented as percentages.

Results

Initially, 5,363 patients were identified with the CPT code 63081 and other inclusion criteria. The exclusion criteria were sequentially applied, excluding 2,451 patients (Fig. 1). Subsequently, analyses included 2,909 patients for prolonged LOS (23.7%), 2,908 patients for non-home discharges (5.5%), and 2,912 patients for major complications (2.8%). The characteristics of the 2,912 patients before outcome-specific exclusions are presented in Table 1.

The model for predicting prolonged LOS, non-home
discharges, and major complications demonstrated mean AUROCs of 0.802, 0.816, and 0.702, respectively. These findings highlight the models’ discriminatory capacities: fair (AUROC >0.7) for differentiating patients with from those without major complications and good (AUROC >0.8) for distinguishing between those with and without prolonged LOS and non-home discharges [26]. A comprehensive analysis of model performance metrics is presented in Table 2.

Figs. 2 and 3 illustrate the ROC curves and PRCs for the three outcomes of interest, respectively. Fig. 4 displays the SHAP bar plots for the models, demonstrating the collective contribution of the top 10 features to the predictions for each outcome. The bar length signifies the mean SHAP value, reflecting an attribute’s influence strength on the predicted outcome, with features organized by significance order from top to bottom. Single-versus multiple-level surgery, age, BMI, preoperative hematocrit, and ASA physical status emerged repetitively as the most important variables for each outcome.

**Discussion**

This study aimed to develop ML models for predicting adverse short-term ACC outcomes and improve their accessibility by creating a web application for healthcare professionals. In this tool, patient informa-
tion is entered to generate personalized risk estimates for each outcome, potentially offering an advantage over traditional methods, often relying on qualitative risk communication or quantitative estimates based on clinicians’ experiences and population statistics. By providing patient-specific predictions, our models can identify patients at increased risk for postoperative adverse events, facilitating more informed discussions of risks during preoperative counseling. In addition, the web application has potential implications for quality assurance by allowing the review of cases with adverse events in patients previously deemed at low risk, potentially uncovering institutional deficiencies and guiding policies to optimize resource allocation and improve patient outcomes.

The approach described herein serves as a robust tool to estimate outcome probabilities for patients undergoing ACC while improving the interpretability of these predictions. As presented in Fig. 4, the SHAP bar charts provide a global perspective on how the models arrive at their predictions overall. Rather than focusing on individual predictions, this holistic view examines pat-
terns, trends, and relationships identified by the models across the entire dataset. On the contrary, SHAP visualizations incorporated into our web application offer granular local interpretations tailored to each patient. This unique feature ensures that users gain clear insight into how predictions relate to particular variable values for a patient. Therefore, our methodology enables an unprecedented level of understanding of the factors that influence outcomes that were not feasible with previous models or platforms.

Although the current web application offers a useful interface to predict adverse ACC outcomes, it is intended as a research tool and should not presently direct clinical decisions. Further validation across varied patient groups and institutions is imperative to confirm its predictive precision. We hope this tool provides an initial foundation for more exhaustive models integrating extra details such as imaging results and granular clinical information to further improve accuracy and clinical utility. As with any predictive method, the generated estimates must be interpreted within the whole clinical context of each patient to enable personalized surgical counseling and planning.

The limitations share similarities with those associated with existing online prognostic models. First, the patient population within the ACS-NSQIP database may not fully reflect real-world patients with ACC. Biases could arise relating to the hospitals because participating institutions potentially possess superior infrastructures and resources to average facilities. In addition, patients in the database may have different health states, age distribution, or socioeconomic backgrounds from the general population. Huffman et al. [27] validated the use of ACS-NSQIP for examining postoperative outcomes, and such factors affect the generalizability of findings despite the overall dependability of the data source. Second, database studies are invariably prone to inaccuracies from coding errors or other misclassifications. Despite the widespread adoption of the ACS-NSQIP, few studies have investigated its coding precision for specific domains. For example, Rolston et al. [28] identified multiple inconsistencies in neurosurgical CPT designations. Finally, we did not establish causal connections between patient attributes and outcomes. Our models are not suited for causal inference and do not provide insight into mechanisms behind the observed predictor–outcome associations. We strongly caution against causal interpretations stemming from our results.

Conclusions

This study has considerably enhanced the prediction of ACC outcomes by implementing advanced ML techniques. A major contribution is the creation of an accessible web application, highlighting the practical value of the developed models. Our findings imply that
ML can serve as an invaluable supplementary tool to stratify patient risk for this procedure and can predict diverse postoperative complications. This methodology has the potential to transform patient counseling discussions toward a more personalized, patient-centered, data-driven approach. Overall, this work signifies a meaningful step forward in enabling precision medicine for spinal procedures.

**Key Points**

- The study demonstrated the effectiveness of machine learning models in predicting outcomes of anterior cervical corpectomy, with models achieving good discriminatory ability in predicting prolonged length of hospital stay and non-home discharges and fair discriminatory ability in predicting major complications.
- Important variables influencing postoperative outcomes included single- versus multiple-level surgery, age, body mass index, preoperative hematocrit levels, and American Society of Anesthesiologists physical status.
- An accessible web application was created for healthcare professionals for the utilization of these machine learning models, prediction of each patient outcome, and assistance in preoperative counseling and risk stratification.
- The study highlights the potential of machine learning to contribute to precision medicine in spinal surgery by enabling person

**Conflict of Interest**

No potential conflict of interest relevant to this article was reported.

**Acknowledgments**

Data for this study were obtained from the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP). ACS NSQIP participant use file access is a benefit of NSQIP participation and is reserved for staff at participating and active ACS NSQIP hospitals. ACS policies do not allow access or sale to nonparticipants. Additional information is provided in [https://www.facs.org/quality-programs/data-and-registries/acs-nsqip/participant-use-data-file/](https://www.facs.org/quality-programs/data-and-registries/acs-nsqip/participant-use-data-file/). The American College of Surgeons National Surgical Quality Improvement Program and the hospitals participating in the American College of Surgeons National Surgical Quality Improvement Program 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. The source code for preprocessing and analyzing the data is available on GitHub ([https://github.com/mertkarabacak/NSQIP-ACC](https://github.com/mertkarabacak/NSQIP-ACC)).

**ORCID**

Mert Karabacak: https://orcid.org/0000-0002-9263-9893; Alexander Schupper: https://orcid.org/0000-0002-3322-8138; Matthew Carr: https://orcid.org/0000-0002-8654-3828; Konstantinos Margetis: https://orcid.org/0000-0002-3715-8093

**Author Contributions**

Conceptualization: MK, AJS, MTC, KM. Methodology: MK, KM. Data curation: MK. Formal analysis: MK. Visualization: MK. Project administration: MK, KM. Writing–original draft preparation: MK. Writing–review and editing: AJS, MTC, KM. Supervision: KM. Final approval of the manuscript: all authors.

**Supplementary Materials**

Supplementary materials can be available from [https://doi.org/10.31616/asj.2024.0048](https://doi.org/10.31616/asj.2024.0048). Supplement 1. CPT codes used for exclusion.

**References**


