Title

Machine learning-driven identification of novel patient factors for prediction of major complications after posterior cervical spinal fusion

Journal

European Spine Journal, 31(8)

ISSN

0940-6719

Authors

Shah, Akash A

Devana, Sai K

Lee, Changhee

et al.

Publication Date

2022-08-01

DOI

10.1007/s00586-021-06961-7

Peer reviewed

HHS Public Access

Author manuscript

Eur Spine J. Author manuscript; available in PMC 2023 August 01.

Published in final edited form as:

Eur Spine J. 2022 August ; 31(8): 1952–1959. doi:10.1007/s00586-021-06961-7.

Machine learning driven identification of novel patient factors for prediction of major complications after posterior cervical spinal fusion

Akash A. Shah[1,†], **Sai K. Devana**[1], **Changhee Lee**[2], **Amador Bugarin**[1], **Elizabeth L. Lord**[1], **Arya N. Shamie**[1], **Don Y. Park**[1], **Mihaela van der Schaar**[2,3], **Nelson F. SooHoo**[1]

[1]Department of Orthopaedic Surgery, David Geffen School of Medicine at UCLA, Los Angeles, CA

[2]Department of Electrical and Computer Engineering, University of California, Los Angeles, CA

[3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom

Abstract

Purpose: Posterior cervical fusion is associated with increased rates of complications and readmission when compared to anterior fusion. Machine learning (ML) models for risk stratification of patients undergoing posterior cervical fusion remain limited. We aim to develop a novel ensemble ML algorithm for prediction of major perioperative complications and readmission after posterior cervical fusion as well as identify factors important to model performance.

Methods: This is a retrospective cohort study of adults who underwent posterior cervical fusion at non-federal California hospitals between 2015–2017. The primary outcome was readmission or major complication. We developed an ensemble model predicting complication risk using an automated ML framework. We compared performance with standard ML models and logistic regression (LR), ranking contribution of included variables to model performance.

Results: Of the included 6,822 patients, 18.8% suffered a major complication or readmission. The ensemble model demonstrated slightly superior predictive performance compared to LR and standard ML models. The most important features to performance include sex, malignancy, pneumonia, stroke, and teaching hospital status. Seven of the ten most important features for the ensemble model were markedly less important for LR.

Conclusion: We report an ensemble ML model for prediction of major complications and readmission after posterior cervical fusion with a modest risk prediction advantage compared

[†] Corresponding author: Akash A. Shah, Department of Orthopaedic Surgery, David Geffen School of Medicine at UCLA, 10833 Le Conte Avenue 76-116 CHS, Los Angeles, CA 90095, aashah@mednet.ucla.edu, Phone: +1 714-331-8707.

Conflicts of interest / competing interests: The authors declare they have no relevant financial interests.

Code availability: The code can be made available from the corresponding author on reasonable request.

Ethics approval: not applicable

Consent to participate: not applicable

Consent for publication: not applicable

to LR and benchmark ML models. Notably, the features most important to the ensemble are markedly different from those for LR, suggesting that advanced ML methods may identify novel prognostic factors for adverse outcomes after posterior cervical fusion.

Keywords

posterior cervical fusion; machine learning; complications; readmission; outcomes

Introduction

Degenerative disease of the cervical spine is common, present in over half of those over 55 years. Cervical spondylotic myelopathy is the most frequent cause of spinal cord dysfunction in this age group.[1] Decompression and fusion is commonly performed to prevent further neurologic deterioration in the setting of myelopathy, as well as to improve functional status.[1,2] Cervical fusion is also performed for treatment of malignancy, trauma, and infection. The rate of cervical fusion is increasing across all ages for all indications; there has been a 30% increase in cervical fusion between 2001 and 2010.[3]

Cervical fusion through a posterior approach may be necessary in the setting of multilevel decompression, a posterior mass compressing the spinal cord (e.g. epidural abscess, malignancy), or posterior element fractures. Posterior cervical fusion is associated with increased length of hospitalization, complication rates, and unplanned readmission rates compared to anterior fusion.[4,5] Furthermore, the average age and comorbidity burden of patients undergoing cervical fusion have increased, elevating the risk of perioperative complications.[4] Given the cost and morbidity associated with perioperative complications and unplanned readmissions, it would be of great utility to improve prediction of which patients are likely to suffer poor outcomes.

AutoPrognosis is a novel ML framework that employs an ensemble of ML methods and traditional statistical approaches, optimizing them into a single well-calibrated ensemble model for outcome prediction across diverse datasets.[6] Furthermore, AutoPrognosis obviates the need for clinicians to choose a specific modeling technique or manually tune hyperparameters. AutoPrognosis has been successfully implemented for prediction of cardiovascular risk in patients with diabetes and survival of cystic fibrosis patients, as well as for major complications after total hip arthroplasty.[7–9]

Using AutoPrognosis, we aim to build a novel ensemble ML algorithm to identify the relative importance of patient factors for prediction of major complication or unplanned readmission after posterior cervical spinal fusion. Secondarily, we aim to compare the performance of this ML ensemble against logistic regression and other standard ML models. We additionally aim to compare the factors most important to the performance of each tested model. We hypothesize that AutoPrognosis will demonstrate superior predictive performance as well as identify novel prognostic features for prediction of adverse outcomes after posterior cervical fusion as compared to logistic regression.

Methods

Study Design and Subjects

This study is a retrospective review utilizing the California Office of Statewide Health and Planning and Development (OSHPD) Patient Discharge Database, a mandatory statewide discharge database containing data for all non-federal hospital admissions in California. Patients in this database are assigned a unique record linkage number that allows patients to be tracked longitudinally for complications and readmissions. We included adults ≥18 years who underwent posterior cervical fusion between 2015–2017 identified using International Classification of Diseases, Tenth Revision (ICD-10) procedure codes for this procedure (Supplementary Table 1).

Outcome and other Variables

The primary outcome measure was any major complication or readmission after index fusion. Complications were identified by adapting codes from performance measures developed by the Centers for Medicare and Medicaid (CMS) for total joint replacement.[10] These complications include acute myocardial infarction, pulmonary embolism, pneumonia, and surgical site bleeding or infection. Myocardial infarction, sepsis, and pneumonia must have occurred during the index admission or within seven days of start of index admission. Pulmonary embolism must have occurred during the index admission or within 30 days of admission. Surgical site bleeding or wound infection must occur during the index admission or within 90 days. The ICD-10 diagnosis and procedure codes used to identify surgical site bleeding and wound complications are specific to cervical spine surgery.

Explanatory features collected for the cohort include patient demographic characteristics (e.g. age, sex, insurance type) and medical comorbidities identified using the CMS Condition Categories as defined by the CMS Hierarchical Condition Category (HCC) risk adjustment model.

Model development and evaluation

AutoPrognosis builds an ensemble model comprised of pipelines. Each pipeline consists of an imputation algorithm, a feature processing algorithm, a classification algorithm, and a calibration method. AutoPrognosis automates the model hyperparameters for each classification method and identifies the best-performing pipelines out of a large space of possible pipelines. Each pipeline is assigned a weight based on the algorithm's posterior belief about each pipeline's utility.[8] In addition to AutoPrognosis, we built five standard ML benchmark models that span different classes of ML modeling approaches: logistic regression (a linear classifier), random forest (a tree-based ensemble classifier), AdaBoost, gradient boosting machines, and XGBoost (boosting ensemble classifiers).[11–14] We implemented logistic regression, random forest, AdaBoost, and gradient boosting machines using the *scikit-learn* Python library.[15] XGBoost was built using the *xgboost* Python library.[14] The classification methods included in AutoPrognosis are shown in Table 1.

We evaluated discrimination and calibration of the prognostic models using five-fold stratified cross-validation. In every cross-validation fold, the training cohort (80% of the study population) was used to derive the AutoPrognosis model and the ML benchmark models. A hold-out testing cohort (20% of the study population) was used for performance evaluation. We report the mean and 95% confidence intervals for all models.

Discrimination was assessed by the area under the receiver operating characteristic curve (AUROC). AUROC represents the probability that a randomly selected patient who experienced an outcome was assigned a higher risk by the model than a patient who did not experience the outcome. Calibration measures the agreement between the model's predictions and observed outcomes in the study population. The Brier score is the mean squared error between observed values and predicted probabilities; it measures both discrimination and calibration. Values close to zero indicate a more accurate model.[16]

The area under the precision-recall curve (AUPRC) is a useful metric when analyzing an imbalanced dataset in which negative cases far outnumber positive cases. The precision-recall curve is constructed by plotting positive predictive value (precision) versus sensitivity (recall). Unlike AUROC, the baseline AUPRC is the proportion of true positive cases. An ideal classifier predicts every positive case (perfect recall) without marking any negative case as positive (perfect precision) and will return an AUPRC of 1. Random prediction will result in the baseline AUPRC. The higher the AUPRC is compared to the baseline AUPRC, the better the model handles positive cases.

Feature importance

We utilize the partial dependence function described by Friedman to measure the importance of an individual feature by assessing the average effect in predicted risks when its value is altered.[13] Specifically, x_c is a chosen target feature in the set of input features χ and $\chi_{\setminus c}$ be its complement, i.e., $\chi = \chi_{\setminus c} \cup x_c$, and $r(\chi) = r(\chi_{\setminus c}, x_c)$ be the predicted risk by our trained model. We then define the feature importance score for an individual feature x_c by averaging $r(\chi_{\setminus c}, x_c = 1) - r(\chi_{\setminus c}, x_c = 0)$ for binary features and $r(\chi_{\setminus c}, x_c = \max(x_c)) - r(\chi_{\setminus c}, x_c = \min(x_c))$ where $\max(x_c)$ and $\min(x_c)$ are the maximum and minimum of feature x_c.

Results

Baseline cohort demographics

A total of 6,822 patients met inclusion criteria for this study. The median age of the cohort was 64 years and the majority of patients (55.2%) were male. The majority of patients were insured through Medicare (51.5%), followed by private insurance (18.5%). The most common comorbidity present in the cohort was diabetes mellitus (12.9%), followed by coronary atherosclerosis (10.8%) and chronic obstructive pulmonary disease (10.2%). Five hundred and eighty-one patients (8.5%) had chronic kidney disease and 538 patients (7.9%) had a history of colorectal or bladder cancer. Over a third of patients (35.2%) had their procedure performed at a teaching hospital. A complete description of cohort demographics is provided in Table 2.

There were 1,279 patients (18.8%) who suffered a major complication or 30-day readmission. The most common perioperative complications were pneumonia, sepsis, and pulmonary embolism (Table 3).

Performance of models

Algorithms predicting risk of major complication or readmission after index fusion were built with AutoPrognosis, logistic regression, and four standard ML benchmark models. The AutoPrognosis model was built with a weighted ensemble of seven ML pipelines (Table 4). This ensemble model demonstrates higher discrimination (AUROC 0.679 ± 0.011) than logistic regression (AUROC 0.651 ± 0.014). It also outperforms the four standard benchmark ML models. It has a Brier score of 0.158 ± 0.001; the remaining models are similarly well-calibrated (Table 5). The receiver-operating characteristic curves for the ensemble and logistic regression models are depicted in Figure 2. The AUPRC of the ensemble model (0.377 ± 0.015) exceeds that of the logistic regression or any benchmark model (Table 6). A random classifier would result in an AUPRC of 0.188, which is the proportion of positive cases in this cohort. The precision-recall curves for the AutoPrognosis and logistic regression models are shown in Figure 3. A confusion matrix for the AutoPrognosis and logistic regression models on the testing with a decision threshold of 0.188 – the proportion of complications in the overall cohort – is provided in Table 7. Random chance is simulated by a biased coin that will randomly classify a patient as positive for complications 18.8% of the time.

Comparison of predictive factors

The relative importance of each variable to model performance for AutoPrognosis and logistic regression are displayed in Table 8. The features most important for risk prediction in AutoPrognosis are colorectal/bladder cancer, history of complications, and aspiration/bacterial pneumonia. The features that are most important for AutoPrognosis differ from those that are most important for logistic regression. Male sex, viral pneumonia, stroke, implant complication, teaching hospital status, circulatory disease, and chronic kidney disease are among the ten most important features for AutoPrognosis model performance but not for logistic regression.

Discussion

As the number of cervical fusions has increased, the age and comorbidity burden of the patients undergoing these surgeries have also increased.[4] Determining which patients are likely to suffer complications would allow for appropriate risk stratification patients prior to surgery. The majority of models for prediction of adverse outcomes after cervical spinal fusion have been developed with logistic regression.[1,4,17–19] ML methods have grown in popularity in recent years due to their ability to detect indirect, nonlinear, and multivariate effects through iterative learning processes with a sensitivity that traditional regression techniques may lack. We report the use of an algorithmic framework that automates generation of an ML-based ensemble model for prediction major perioperative complications and unplanned readmission after posterior cervical fusion. AutoPrognosis automates model selection and hyperparameter tuning, facilitating the use of optimized ML

pipelines in clinical prognostic research by investigators who may not possess expertise in advanced ML methods. The reported ensemble model has slightly improved discrimination in terms of both AUROC and AUPRC compared to logistic regression. More importantly, the ensemble model identifies novel prognostic features important for model performance that differ dramatically from those important for logistic regression.

The most important binary features for performance of the ensemble model include history of medical complications, colorectal/bladder cancer, and aspiration pneumonia. History of complications in past hospital admissions (e.g. excessive transfusion requirement, implant infection) is intuitively contributory to developing future complications. Patients with a history of malignancy may be more likely to have a higher comorbidity burden and thus more likely to suffer from perioperative complications; malignancy has been shown to be a risk factor for post-operative complications in the elective spine and arthroplasty literature.[9,20] Patients with a history of aspiration pneumonia are more likely to be of advanced age and lower functional status, both risk factors for poor outcomes after posterior cervical fusion.[4,18,21] The finding that the number of CMS Condition Categories comorbidities is the most important continuous variable to model performance is in line with studies showing increased comorbidity burden as being associated with major complications and readmissions after cervical fusion.[4,20]

Notably, seven of the top ten features most important to the AutoPrognosis model were markedly less important for logistic regression: sex, viral pneumonia, stroke, implant complications, teaching hospital status, circulatory disease, and chronic kidney disease. Prior stroke, pneumonia, chronic kidney disease, and peripheral vascular disease are markers of a patient with lower baseline physiologic reserve who may be at increased risk of perioperative complications. While history of stroke has been shown to be associated with post-operative delirium after cervical spine surgery, it has not been specifically implicated in influencing complication risk after posterior cervical fusion.[22] Patients who suffered a stroke may have decreased functional status – a risk factor for complications after posterior cervical fusion.[18] Although peripheral vascular disease and chronic renal disease are the 9th and 10th most important features for the ensemble model, they are markedly less important in logistic regression (26th and 56th, respectively); they have both been identified as risk factors for poor outcomes and readmission after posterior cervical fusion.[4,17]

Patient sex is the third most important feature for the ensemble model but the 27th most important for logistic regression. Male sex has been shown to be a predictor for post-operative complications after major orthopaedic, thoracic, urologic, and neurosurgical procedures; it has also been specifically implicated as a risk factor for poor outcomes after posterior cervical fusion.[4,23] This difference in outcomes between male and female patients has been hypothesized to stem from a higher overall comorbidity burden in males, higher rates of smoking of alcohol use among males, and lower pre-operative functional status of male patients compared to female patients.[23] Finally, teaching hospital status is the 8th most important feature to performance of the ensemble model but the 51st most important for logistic regression. Resident participation is associated with increased likelihood of blood transfusion, increased operative time, and prolonged length of stay after elective posterior cervical fusion.[24] Teaching hospitals also have the responsibility of educating

nurses, operating room technicians, medical students, and other health care provider trainees that may incrementally contribute to increased complication risk.

The finding that the features most important for AutoPrognosis and logistic regression dramatically differ is significant but must be carefully interpreted. Taken together, the findings that AutoPrognosis has slightly superior predictive performance to logistic regression and that chronic kidney disease is more important to AutoPrognosis than it is for logistic regression do not necessarily imply that this feature is more strongly associated with poor outcomes after posterior cervical fusion than the literature has reported. No conclusion regarding correlation between feature and outcome can be drawn since AutoPrognosis is designed for classification problems and not for statistical inference. The appropriate conclusion is that the ensemble model and logistic regression treat the same predictive features differently. The performance of AutoPrognosis stems from its ability to detect complex non-linear relationships between variables that traditional regression techniques are unable to capture.

This study has multiple limitations. The retrospective use of a de-identified state database without access to patient charts limits the granularity of extracted features and outcomes. Reliance on diagnosis codes in an administrative database may underestimate complication rates compared to chart review. Furthermore, ICD-10 procedure codes are often unreliable for diagnosis and do not allow determination of the indication for cervical fusion. Additionally, this database does not contain patient-reported outcomes. It should be noted that selection bias exists in this cohort as it is comprised of patients for whom it was determined that the perceived benefit of surgery outweighed the risks. While we aimed to capture only complications associated with surgery by limiting inclusion of complications to the immediate perioperative period, we cannot exclude the possibility that a small number of complications that occurred in the perioperative period are unrelated to posterior cervical fusion. Unfortunately, spine-specific complications (e.g. implant-related complications, neurologic complications, revision surgery) are not available in this dataset; diagnosis codes specific to these complications that we can effective query are similarly not available. Future application of this ensemble method to a more granular institutional database may allow for identification and prediction of these complications. Finally, a strength of advanced ML techniques is their ability to magnify nuances of a dataset; unfortunately, this may have the unintended effect of exacerbating biases present in the data. Disparities in data collection and care between different groups in the training cohort may be amplified, potentially causing harm to underrepresented groups such as those of lower socioeconomic status and ethnic minorities.[25] Future studies with multi-institutional or prospective designs are thus necessary.

With a sample of 6,822 patients, we report a novel ensemble ML algorithm that predicts major perioperative complications and readmission after posterior cervical fusion – major drivers of morbidity and mortality. This model is well-calibrated and displays a small advantage in predictive accuracy compared to regression and standard benchmark ML models. Although the discrimination of this model is fair, an AUROC value below 0.70 is not robust enough for the algorithm to be deployed as a risk calculator to guide clinical decisions at this time. The key finding of this study, however, is that the ensemble model

identified novel prognostic features that were different from those most important for logistic regression. This suggests that ensemble methods like the one we describe may be able to uncover interactions between features that are not readily detected by logistic regression. Further application of this novel ensemble method to a high-quality prospective cohort would be of great utility and allow for the development of an ML-driven tool to improve prediction of adverse outcomes after posterior cervical fusion.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: The research reported in this publication was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under the Ruth L. Kirschstein National Research Service Award Number T32AR059033.

References

1. Veeravagu A, Connolly I, Lamsam L, et al. [2016] Surgical outcomes of cervical spondylotic myelopathy: an analysis of a national, administrative, longitudinal database. Neurosurg Focus 40(6):E11.
2. Fehlings M, Wilson J, Kopjar B, et al. [2013] Efficacy and safety of surgical decompression in patients with cervical spondylotic myelopathy: results of the AOSpine North America prospective multi-center study. J Bone Jt Surg Am 95(18):1651–1658.
3. Goz V, Weinreb J, McCarthy I, Schwab F, Lafage V, Errico T [2013] Perioperative complications and mortality after spinal fusions: analysis of trends and risk factors. Spine (Phila Pa 1976) 38(22):1970–1976. [PubMed: 23928714]
4. Memtsoudis S, Hughes A, Ma Y, Chiu Y, Sama A, Girardi F [2011] Increased in-hospital complications after primary posterior versus primary anterior cervical fusion. Clin Orthop Relat Res 469:649–657. [PubMed: 20838946]
5. Shamji M, Cook C, Pietrobon R, Tackett S, Brown C, Isaacs R [2009] Impact of surgical approach on complications and resource utilization of cervical spine fusion: a nationwide perspective to the surgical treatment of diffuse cervical spondylosis. Spine J 9(1):31–38. [PubMed: 18790678]
6. Alaa A, van der Schaar M [2018] AutoPrognosis: automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. Proc 35th Int Conf Mach Learn PMLR 80:139–148.
7. Alaa AM, Bolton T, Angelantonio E Di, Rudd JHF, van der Schaar M [2019] Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PLoS One 14(5):1–17.
8. Alaa AM, van der Schaar M [2018] Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. Sci Rep 8(1):1–19. doi:10.1038/s41598-018-29523-2 [PubMed: 29311619]
9. Shah AA, Devana SK, Lee C, Kianian R, van der Schaar M, SooHoo NF [2021] Development of a novel, potentially universal machine learning algorithm for prediction of complications after total hip arthroplasty. J Arthroplasty 36(5):1655–1662. [PubMed: 33478891]
10. Yale New Haven Health Services Corporation / Center for Outcomes Research & Evaluation. 2017 Procedure-specific measure updates and specifications report hospital-level risk-standardized complication measure: elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) - version 6.0 2017.
11. Breiman L [2001] Random forests. Mach Learn 45:5–32.
12. Ratsch G, Onoda T, Muller K [2001] Soft margins for AdaBoost. Mach Learn. 42:287–320.

13. Friedman JH [2001] Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. doi:10.1214/aos/1013203451
14. Chen T, Guestrin C [2016] XGBoost: a scalable tree boosting system. KDD ‘16 Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min 785–794.
15. Pedregosa F, Varoquaux G, Gramfort A, et al. [2011] Scikit-learn: machine learning in Python. J Mach Learn Res 12(85):2825–2830.
16. Manning DW, Edelstein AI, Alvi HM [2016] Risk prediction tools for hip and knee arthroplasty. J Am Acad Orthop Surg 24(1):19–27. doi:10.5435/JAAOS-D-15-00072 [PubMed: 26604220]
17. Choy W, Lam S, Smith Z, Dahdaleh N [2018] Predictors of 30-day hospital readmission after posterior cervical fusion in 3401 patients. Spine (Phila Pa 1976) 43(5):356–363. [PubMed: 26872307]
18. DePasse J, Durand W, Eltorai A, Palumbo M, Daniels A [2018] Timing of complications following posterior cervical fusion. J Orthop 15(2):522–526. [PubMed: 29881186]
19. Vonck C, Tanenbaum J, Bomberger T, et al. [2018] Short-term outcomes following posterior cervical fusion among octogenarians with cervical spondylotic myelopathy: a NSQIP database analysis. Spine J 18(9):1603–1611. [PubMed: 29454135]
20. Katz A, Mancini N, Karukonda T, Cote M, Moss I [2019] Comparative and predictor analysis of 30-day readmission, reoperation, and morbidity in patients undergoing multilevel ACDF versus single and multilevel ACCF using the ACS-NSQIP dataset. Spine (Phila Pa 1976) 44(23):E1379–E1387. [PubMed: 31348176]
21. Wu C, Chen Y, Wang M, Pinelis E [2017] National trends in admission for aspiration pneumonia in the United States, 2002–2012. Ann Am Thorac Soc 14(6):874–879. [PubMed: 28272915]
22. Radcliff K, Ong K, Lovald S, Lau E, Kurd M [2017] Cervical spine surgery complications and risks in the elderly. Spine (Phila Pa 1976) 42(6):E347–E354. [PubMed: 28291765]
23. Al-Taki M, Sukkarieh H, Hoballah J, et al. [2018] Effect of gender on postoperative morbidity and mortality outcomes: a retrospective cohort study. Am Surg 84(3):377–86. [PubMed: 29559052]
24. Lee N, Kothari P, Kim C, et al. [2018] The impact of resident involvement in elective posterior cervical fusion. Spine (Phila Pa 1976) 43(5):316–323. [PubMed: 26839988]
25. Gianfrancesco M, Tamang S, Yazdany J, Schmajuk G [2018] Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178(11):1544–1547. [PubMed: 30128552]

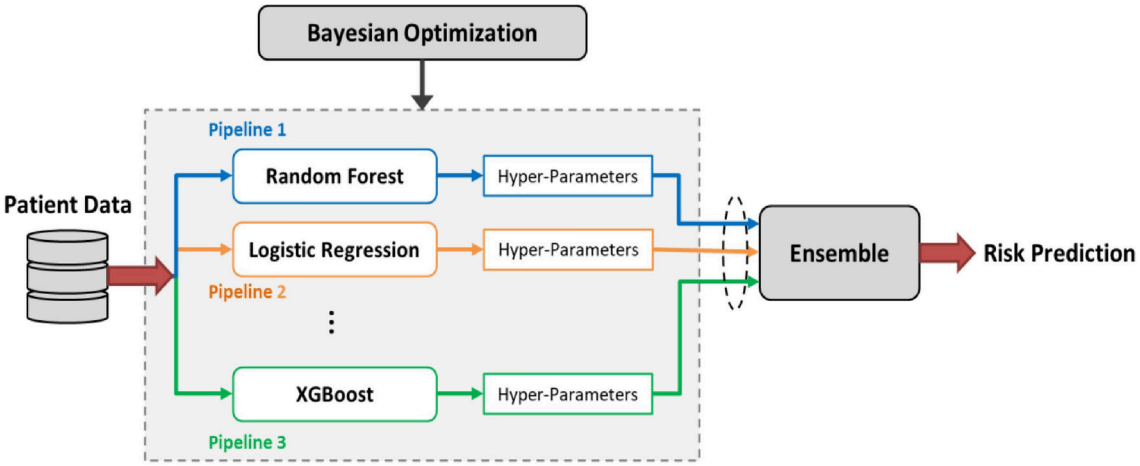

Fig. 1.
Schematic representation of AutoPrognosis workflow

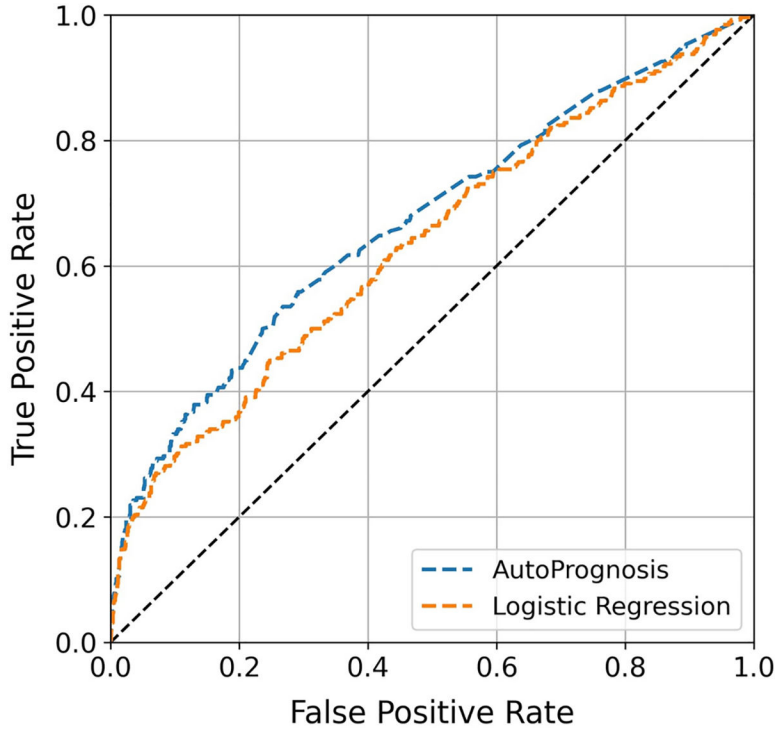

Fig. 2.
Receiver operating characteristic curves for AutoPrognosis and logistic regression

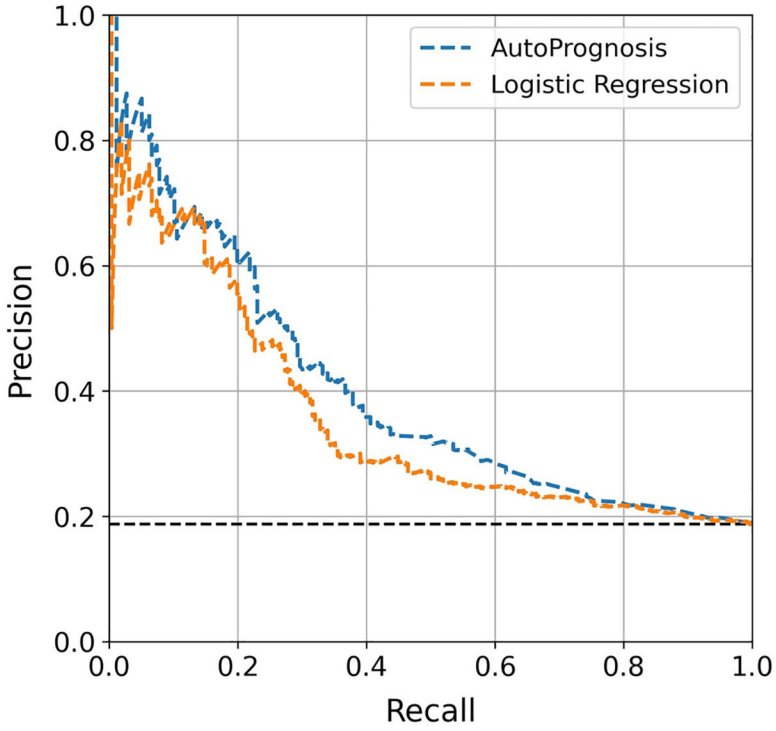

Fig. 3.
Precision-recall curves for AutoPrognosis and logistic regression

Table 1.

List of classification methods included in AutoPrognosis

Classification Methods		
Logistic Regression	Random Forest	Gradient Boosting
XGBoost	AdaBoost	Bagging
Bernoulli NB	Gaussian NB	Multinomial NB
Perceptron	Decision Trees	SVM
LDA	QDA	kNN
Neural Networks		

Table 2.

Baseline cohort demographics

Variable	All Patients (n = 6,822)
Demographics	
	Median (IQR)
Age (years)	64 (55 – 72)
Hospital volume †	305 (143 – 457)
	Number (%)
Male	3,767 (55.22)
Race	
White	4,805 (70.43)
Black	640 (9.38)
Asian / Pacific Islander	476 (6.98)
Native American	34 (0.50)
Other	780 (11.43)
Unknown	87 (1.28)
Insurance	
Medicare	3,512 (51.48)
Private	1,262 (18.50)
Medi-Cal	921 (13.50)
Workers' compensation	261 (3.83)
Other	696 (10.20)
Procedure performed at teaching hospital	2,403 (35.22)
Medical comorbidities	
Diabetes mellitus	878 (12.87)
Coronary atherosclerosis	739 (10.83)
COPD	697 (10.22)
End-stage chronic kidney disease	581 (8.52)
Colorectal/bladder cancer	538 (7.89)
Lung/other cancer	557 (8.16)
Protein-calorie malnutrition	611 (8.96)
Viral or unspecified pneumonia	609 (8.93)
Bacterial or aspiration pneumonia	569 (8.34)
Dementia	582 (8.53)
Major depressive or bipolar disorder	627 (9.19)
Ischemic or unspecified stroke	544 (7.97)
Vertebral fractures without spinal cord injury	714 (10.47)
Spinal cord injury	745 (10.92)
Peripheral vascular disease	632 (9.26)
Implant complications	584 (8.56)
History of prior complications	586 (8.59)
	Mean

Variable	All Patients (n = 6,822)
Number of comorbidities	0.79

IQR = Interquartile range; COPD = chronic obstructive pulmonary disease

† Cases of cervical fusions performed between 2015 and 2017

Table 3.

Major complications and readmission

Complications	All Patients (n = 6,822)
	Number (%)
At least one complication or readmission	1,279 (18.75)
Readmission within 30 days	834 (12.23)
Pneumonia	421 (6.17)
Sepsis	286 (4.19)
Pulmonary embolism	95 (1.39)
Acute myocardial infarction	40 (0.59)
Surgical site bleeding or infection	6 (0.08)

Table 4.

List of the 7 pipelines fitted to the posterior cervical fusion cohort

Pipeline #	Methods	Hyper-Parameters	Weight
1	Gradient Boosting	(learning rate = 0.052, max depth = 2, estimators = 121)	0.315
2	Gradient Boosting	(learning rate = 0.045, max depth = 2, estimators = 122)	0.216
3	Gradient Boosting	(learning rate = 0.042, max depth = 3, estimators = 122)	0.170
4	Gradient Boosting	(learning rate = 0.046, max depth = 3, estimators = 126)	0.091
5	XGBoost	(learning rate = 0.0242, max depth = 2, estimators = 448)	0.091
6	Gradient Boosting	(learning rate = 0.0561, max depth = 3, estimators = 125)	0.062
7	Gradient Boosting	(learning rate = 0.029, max depth = 3, estimators = 122	0.054

Table 5.

Discrimination and calibration of AutoPrognosis, logistic regression, and ML benchmark models

Model	AUROC	Brier score
AutoPrognosis	**0.679 ± 0.011**	0.158 ± 0.037
Logistic Regression	0.651 ± 0.014	0.141 ± 0.001
Gradient Boosting	0.676 ± 0.013	0.139 ± 0.002
XGBoost	0.674 ± 0.015	0.140 ± 0.003
AdaBoost	0.671 ± 0.015	0.248 ± 0.001
Random Forest	0.647 ± 0.011	0.158 ± 0.037

Table 6.

Area under the precision-recall curve for AutoPrognosis, logistic regression, and ML benchmark models

Model	AUPRC
AutoPrognosis	**0.377 ± 0.015**
Logistic Regression	0.362 ± 0.012
Gradient Boosting	0.376 ± 0.017
XGBoost	0.374 ± 0.027
AdaBoost	0.365 ± 0.024
Random Forest	0.298 ± 0.008

Table 7.

Confusion matrix for AutoPrognosis and logistic regression model performance on testing cohort at a decision threshold of 0.188

N = 1,364	Predicted negative	Predicted positive
Observed negative	AP: 853 LR: 833 Chance: 896	AP: 255 LR: 275 Chance: 212
Observed positive	AP:125 LR: 131 Chance: 200	AP:131 LR: 125 Chance: 56

Table 8.

Relative feature importance for AutoPrognosis and logistic regression

Feature	Rank in AutoPrognosis (Rank in logistic regression)	Change to risk prediction
Binary features		
Colorectal/bladder cancer	1 (1)	0.0726
Viral/unspecified pneumonia	2 (11)	0.0328
Male sex	3 (27)	0.0277
History of prior complications	4 (9)	0.0201
Aspiration/bacterial pneumonia	5 (3)	0.0181
Stroke	6 (12)	0.0103
Implant complication	7 (30)	0.0082
Teaching hospital	8 (51)	0.0074
Peripheral vascular disease	9 (26)	0.0070
Chronic kidney disease	10 (56)	0.0029
Continuous features		
Number of comorbidities	1 (1)	0.1637
Hospital volume	2 (2)	−0.0234
Age	3 (3)	−0.0080
Insurance status		
Medicare	Reference	0
Other	1 (2)	0.0453
Medi-Cal	2 (4)	0.0300
Private	3 (3)	−0.0239
Workers' compensation	4 (1)	−0.0175