2019

Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms

Farrokh Farrokhi

Quinlan D. Buchlak

Matt Sikora

Nazanin Esmaili

Maria Marsans

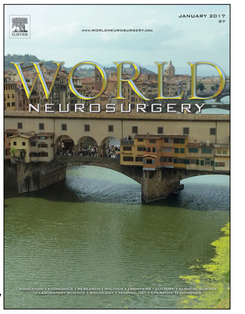
See next page for additional authors

This article was originally published as:
Farrokhi, F., Buchlak, Q. D., Sikora, M., Esmaili, N., Marsans, M., McLeod, P., Mark, J., Cox, E., Bennett, C., & Carlson, J. (2019). Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms. *World Neurosurgery, Early View, Online First*.

Original article available here:
https://doi.org/10.1016/j.wneu.2019.10.063

Authors
Farrokh Farrokhi, Quinlan D. Buchlak, Matt Sikora, Nazanin Esmaili, Maria Marsans, Pamela McLeod, Jamie Mark, Emily Cox, Christine Bennett, and Jonathan Carlson

Journal Pre-proof



Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms

Farrokh Farrokhi, MD, Quinlan D. Buchlak, MPsych MBIS, Matt Sikora, BA, Nazanin Esmaili, PhD, Maria Marsans, PAC, Pamela McLeod, PAC, Jamie Mark, ARNP, Emily Cox, PhD, Christine Bennett, MBBS, Jonathan Carlson, MD PhD

PII: S1878-8750(19)32681-6

DOI: https://doi.org/10.1016/j.wneu.2019.10.063

Reference: WNEU 13534

To appear in: *World Neurosurgery*

Received Date: 6 July 2019

Revised Date: 9 October 2019

Accepted Date: 10 October 2019

Please cite this article as: Farrokhi F, Buchlak QD, Sikora M, Esmaili N, Marsans M, McLeod P, Mark J, Cox E, Bennett C, Carlson J, Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms, *World Neurosurgery* (2019), doi: https://doi.org/10.1016/j.wneu.2019.10.063.

Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms

Farrokh Farrokhi MD[1], Quinlan D. Buchlak MPsych MBIS[2], Matt Sikora BA[1], Nazanin Esmaili PhD[2,3,4], Maria Marsans PAC[1], Pamela McLeod PAC[5], Jamie Mark ARNP[6], Emily Cox PhD[7], Christine Bennett MBBS[2], Jonathan Carlson MD PhD[5]

[1]Neuroscience Institute, Virginia Mason Medical Center, Seattle, WA, USA
[2]School of Medicine, University of Notre Dame Australia, Sydney, NSW, Australia
[3]Department of Medicine, University of Toronto, Toronto, ON, Canada
[4]Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW, Australia
[5]Inland Neurosurgery and Spine Associates, Spokane, WA, USA
[6]Selkirk Neurology, Spokane, WA, USA
[7]Providence Medical Research Center, Providence Health & Services, Spokane, WA, USA

Corresponding author:
Quinlan Buchlak
The University of Notre Dame
160 Oxford St
Sydney, Australia, 2015
E-mail: quinlan.buchlak1@my.nd.edu.au

Element counts: Abstract word count = 248; text word count (excluding abstract) = 4191; number of references = 87; number of tables = 5; number of figures = 4.
Disclosures: All authors have reviewed and approved this manuscript and have no relevant financial or other conflicts of interest with regard to this research and its publication.
Keywords: Risk stratification, data imputation, machine learning, supervised learning, gradient boosting machines, deep brain stimulation, neurosurgery

ABSTRACT

Background: Deep brain stimulation (DBS) surgery is an option for patients experiencing medically resistant neurological symptoms. DBS complications are rare; finding significant predictors requires a large number of surgeries. Machine learning algorithms may be used to effectively predict these outcomes. The aims of this study were to (1) investigate preoperative clinical risk factors, and (2) build machine learning models to predict adverse outcomes.

Methods: This multicenter registry collected clinical and demographic characteristics of patients undergoing DBS surgery (n=501) and tabulated occurrence of complications. Logistic regression was used to evaluate risk factors. Supervised learning algorithms were trained and validated on 70% and 30%, respectively, of both oversampled and original registry data. Performance was evaluated using area under the receiver operating characteristics curve (AUC), sensitivity, specificity and accuracy.

Results: Logistic regression showed that the risk of complication was related to the operating institution in which the surgery was performed (OR=0.44, confidence interval [CI]=0.25-0.78), BMI (OR=0.94,CI=0.89-0.99) and diabetes (OR=2.33,CI=1.18-4.60). Patients with diabetes were almost three times more likely to return to the operating room (OR=2.78,CI=1.31-5.88). Patients with a history of smoking were four times more likely to experience postoperative infection (OR=4.20,CI=1.21-14.61). Supervised learning algorithms demonstrated high discrimination performance when predicting any complication (AUC=0.86), a complication within 12 months (AUC=0.91), return to the operating room (AUC=0.88) and infection (AUC=0.97). Age, BMI, procedure side, gender and a diagnosis of Parkinson’s disease were influential features.

Conclusions: Multiple significant complication risk factors were identified and supervised learning algorithms effectively predicted adverse outcomes in DBS surgery.

INTRODUCTION

The primary aim of this study was to look at which preoperative clinical factors were related to complications that develop in deep brain stimulation (DBS) therapy. DBS is a safe, effective and common surgical intervention for a range of neurological disorders including Parkinson's disease and essential tremor [1–7]. Through electrodes implanted in the brain, DBS therapy stimulates deep subcortical brain structures, including the subthalamic nucleus (STN), the ventral intermedius nucleus (VIM) and the globus pallidus (GPi) to alleviate neurological symptoms like tremor, motor fluctuations, and rigidity [4,5,8]. It is a treatment modality that is considered when a patient's symptoms have not been satisfactorily alleviated by medical management [9–14].

DBS therapy requires an initial electrode implantation operation and subsequent surgery to place device generators. Potential complications arising from DBS surgery include infection, intracerebral hemorrhage, seizures and hardware failure, which can lead to unplanned return to the operating room. Post-operative readmission rates range from 1.9% (30-day) to 4.3% (90-day) [1]. Factors likely associated with complications include age, smoking history, obesity, diabetes, hypertension and facility surgical volume [1,15]. Advanced age and hypertension have been associated with the risk of intracranial hemorrhage [16], and readmission after DBS surgery has been associated with preoperative coronary artery disease, obesity and a history of smoking [1]. Further, there is a seasonal variation in DBS infection, often referred to as the July effect [17].

Integrating preoperative risk assessment into standard clinical care fosters a shared decision making process between the surgical team, the patient and clinical enablers [18]. Performing pre-operative risk assessment for DBS procedures is challenging due to limited data suggesting the contributions of individual risk factors to post-operative complications. It is arguable that the literature surrounding DBS surgery risk remains inconclusive because the low frequency of complications limits the power and sensitivity of traditional statistical methods. To study this problem, a multi-institutional database of complications and risk factors was compiled, and a pilot study analysed it. Similar to the literature, the only relationship found was an association between smoking and infection risk. The standard statistical methods applied were ineffective at determining significant clinical risk factors related to complications, such as body mass index, diabetes, hypertension, smoking, and age. So, a different approach to identifying relationships between complications and risk factors, involving the use of machine learning, was designed and deployed.

Machine learning, a branch of artificial intelligence, represents a powerful set of technologies that enable three main tasks: classification, regression and clustering [19]. Supervised learning involves training algorithms with datasets that contain labelled outcomes for each case. Supervised learning (i.e., classification and regression) uses input features (X) to predict a defined outcome (Y), while unsupervised learning (i.e., clustering) involves analyzing input variables (X) to elucidate patterns and structure in the data. Supervised learning algorithms can predict rare events such as surgical complications [20] and have the potential to improve patient risk stratification, clinical decision making, informed consent and health service planning [18,21–25]. Supervised learning has been used in DBS surgery to predict clinical outcomes [26,27], surgical targets [28,29], side effects [30], discharge status [31] and neurophysiological detection of DBS structures [32–34].

Extreme gradient boosting machines (XGBM) are a type of supervised learning algorithm. It uses decision tree-based learning and shows strong performance on a diverse array of problems. It operates by strategically combining networks of sequential decision trees. Later decision tree models correct inaccuracies in previous models to improve prediction performance [35]. An XGBM model is comprised of an ensemble of decision trees. The development of algorithms that incorporate gradient boosting has produced highly robust regression and classification methods [36]. XGBMs appear to have performed well in various domains [35,37–41] and have been shown to perform particularly well on datasets characterized by class imbalance [42,43]. Many supervised learning algorithms perform well as predictive tools partly because they can estimate complex nonlinear relationships in high volume datasets using weighted statistical functions in a way that cannot be perceived by linear models or clinicians [44–47]. Logistic regression is one such linear classification model. Two advantages it affords are that it is easily interpretable, and it delivers measures of statistical significance.

Class imbalance describes a situation where the number of one event type (e.g., postoperative DBS complications) is very low compared to another event type (e.g., no postoperative DBS complications) [48]. A class is a subcategory within a variable in a dataset. For example, within a variable capturing data on postoperative complications, one class may represent the complication state, while another class may represent the no-complication state. Class imbalance essentially refers to differences in class probabilities [49]. Postoperative DBS complications are low probability events. There is a much higher probability that DBS patients will experience no postoperative complications. This imbalanced outcome probability is what is meant by researchers referring to imbalanced classes. Chawla (2010) states that a dataset is imbalanced if the classification categories are not equally represented [50]. The performance of some supervised learning algorithms is undermined by class imbalance, resulting in output classifications that default simply to the majority class [51–53] . However, class imbalance characterizes many real-world datasets from biomedicine [52], to finance [54], aviation [55] and geoscience [56]. Class imbalance is one of the main barriers to effectively predicting postoperative complications in neurosurgery [18,19,24,31].

Because the class imbalance problem is so prevalent [53], much research in the fields of predictive analytics, data mining and machine learning has focused on developing and testing methods to effectively address it, at both the algorithm and data levels [49,51,52,57–59]. The Synthetic Minority Oversampling Technique (SMOTE) has emerged as one effective method of addressing the class imbalance issue at the data level [59]. It operates by creating additional synthetic cases based upon existing minority cases and the k-nearest neighbor algorithm. It balances the class distribution by synthesizing new additional minority class examples through a process of interpolating between multiple minority class examples that lie together in multidimensional space. In this way, SMOTE has been intentionally designed to avoid the predictive analytics problem of overfitting [53]. Another strength of employing the SMOTE method is that no cases in the dataset need to be excluded from the predictive analysis, which is particularly useful in neurosurgery where cases are not common and datasets are often not large (i.e., hundreds of cases rather than thousands). The application of SMOTE may effectively facilitate the prediction of DBS complications, which would be of substantial utility to clinicians.

This study sought to answer the following two research questions. Can multivariate logistic regression detect significant associations between preoperative variables and postoperative outcomes? Can DBS complications be accurately predicted by applying the XGBM algorithm and SMOTE?

METHOD

Subjects

This study was approved by Institutional Review Boards at each study site. Due to the retrospective nature of the study, the requirement for informed consent was waived. A combined registry was created comprising 501 adults who underwent initial DBS implantation surgeries between October 1997 and May 2018 at two private practices. Procedures included were performed by five neurosurgeons at two neurosurgical centers over a 22-year period. Patients underwent DBS implantation for Parkinson's disease (n=348), essential tremor (n=129), dystonia (n=11) and other indications (n=13).

Surgical Technique

The general surgical technique was relatively similar among all surgeons. Primary surgeons at each institution each had >15 years of experience in DBS surgery. A frame-based approach was used in patients with DBS lead placement (unilateral or bilateral) using Medtronic 3389 or 3379 leads. Microelectrode recording was used in all cases. A single microelectrode was used to identify and confirm the target in all cases. The average number of microelectrode passes per lead was 1.4. Intraoperative imaging of lead location with cone beam CT was performed in some cases beginning in 2008. The majority of patients underwent intraoperative bipolar review of clinical efficacy and side effects in an "awake" state [60]. Generator placement was staged one to two weeks after initial lead implantation. All patients underwent postoperative MRI and/or CT scans within a week of lead implantation.

Data

Pre-existing quality assurance databases of DBS patients and their outcomes from both research sites were combined. Additional retrospective data were collected from electronic medical records. Potential risk factors were recorded, including age, gender, BMI, clinical diagnosis, smoking history, immunosuppression, hypertension (medications taken within 90 days of surgery), diagnosis of diabetes mellitus, hypertension, surgical target (VIM, STN, GPi) and procedure side (left, right, bilateral).

Complication categories were intracranial hemorrhage, readmission, ischemic infarction, seizure, lead fracture, electrode migration, loose or flipping battery needing surgical revision, device malfunction, return to the operating room and infection. An infection was defined as an event requiring surgical removal of hardware, regardless of the time after implantation. This included perioperative infections within 3 months of surgery, as well as delayed infection associated with hardware erosion or other systemic infections and infections after generator replacement surgery that could have been years later. Intracranial hemorrhage was defined as any form of new post-operative bleeding on radiology report, with or without neurological sequelae and not necessarily requiring surgical intervention. Return to the operating room included all surgeries that required a return to the operating room, regardless of time since lead implantation, for indications including haemorrhage, infection, erosion of hardware, fracture of hardware detected on imaging or as open circuit on programming, revision of lead location, revision of flipping or loose generator, or tight extension wires. Four primary outcomes were recorded for each patient: any postoperative complication, a complication within 12 months of surgery, return to the operating room and infection.

Analysis

Unilateral (n=151), simultaneous bilateral (n=296) and staged bilateral lead implantation (n=54) were counted each as a single case. Descriptive statistics, multivariate logistic regression and supervised learning model development were performed using Python 3.6.

Neural network development for BMI data imputation

Missing data can create problems for some supervised learning algorithms and may necessitate dropping entire cases. Further, missing data can adversely affect the validity of results [61]. Out of 501 cases, there were 51 missing BMI values. Given the scarce nature of DBS case data and the resources required to collect it, the research team was motivated to retain as many cases as possible for analysis. Data imputation addresses this issue and various methods can be used [61–65]. Four neural network regression models were developed to impute BMI for the cases with missing data. BMI values were imputed using all pre- and post-operative variables in the dataset. One neural network was selected for imputation regression because it demonstrated the best performance. Mean BMI before and after imputation did not differ significantly (27.57, SD=5.06 and 27.39, SD=5.98, respectively; p=0.58).

Feature selection

Three criteria were used when selecting input features for the models: (1) existing evidence in the literature suggesting a relationship between the feature and the outcome, (2) availability of the feature in the dataset; and (3) clinical expert approval that the feature under consideration was clinically related to the outcome variable.

Multivariate logistic regression to detect associations

Multivariate logistic regression was conducted using the statsmodels [53] and scikitlearn [54] packages. Multivariate model performance, odds ratios (OR) and confidence intervals (CI) were calculated for each risk factor. Features with negligible statistical contribution to multivariate models (z-score <0.02) were excluded and models were subsequently retrained.

XGBM model development for postoperative complication prediction

Multiple classifiers were tested and compared to predict postoperative complication outcomes, including logistic regression, random forests, decision trees and support vector machines. Algorithm performance statistics were compared using multiple metrics including area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, positive predictive value and negative predictive value. XGBM was among the highest performing classifiers. Because of this and previous literature demonstrating strong performance on imbalanced datasets, multiple XGBM models were developed using the XGBoost [66] package. For each of the four primary outcome variables, three XGBM models were created: one using the original dataset, one using the SMOTE dataset [59], and one using a SMOTE training dataset with a non-SMOTE validation dataset.

Each model was trained on a 70% sample of the dataset and validated on the remaining 30%. In the original dataset, this resulted in 350 training cases and 151 validation cases. In the SMOTE oversampled datasets, ratios of training:validation case numbers were as follows: any complication, 585:251; complication within 12 months, 618:266; return to the operating room, 627:269; and infection, 663:285. SMOTE was selected over other techniques to address class imbalance because (1) it allowed retention and use of all cases in the DBS dataset, (2) it was designed to avoid overfitting, and (3) it has been implemented as an accessible Python package.

Hyperparameter tuning involving grid-search with 5-fold cross-validation was used to find optimal XGBM parameters. Grid-search employed 1512 hyperparameter combinations, resulting in 7560 fit cycles for each of the XGBM models. Using the optimal hyperparameters found in the grid-search process, internal cross-validation was conducted with the number of boosting rounds set at 50 and the number of early stopping rounds set at 10. AUC was used as the performance metric in this process.

Predictions were made using the optimized model and the validation test sets. Confusion matrices and performance statistics were computed. Performance metrics included AUC, accuracy, sensitivity, specificity, positive predictive value and negative predictive value [44,67,68]. Feature importance was calculated, decision trees were visualized and receiver operating characteristics (ROC) curves developed. Figure 1 outlines the analysis process overall.

RESULTS

Descriptive statistics are displayed in Table 1. Mean age at implant was 64±10.3 years. The majority of patients were male (63%), were diagnosed with Parkinson's disease (70%), had a BMI of 25 or more (67%) and underwent a simultaneous bilateral (59%) STN procedure (70%). Patient characteristics did not differ significantly between institutions.

Complication Rates

There were 27 (5.4%) infections over the period of observation (mean 455 days). These infections were either perioperative, occurring within 3 months of lead implantation in 13 (2.6%) patients, or delayed in 14 (2.8%) patients. The median time to onset of all infections was 3.3 months. Delayed infections were typically related to hardware erosion, systemic infections, generator replacement, or appeared spontaneously.

Surgical revision of hardware occurred in 26 (5.2%) patients, on average 28 months after initial implantation. These revisions were for lead or extension wire fracture in 18 (3.6%) patients, loose hardware in seven (1.4%), or repositioned leads due to side effects or poor efficacy in eight (1.6%).

Intracranial hemorrhage occurred in 15 (3.0%) patients, all associated with lead implantation. This included intraparenchymal hemorrhage along the lead and subdural hematoma. No deaths occurred in any of these cases. Of these hemorrhages, 2 of 501 patients (0.4%) had substantial morbidity requiring surgical intervention. Other hemorrhages, 13 (2.6%), were observed on imaging, and resolved without surgical treatment or neurological sequelae.

Risk factors identified using logistic regression

Logistic regression demonstrated statistically significant relationships between risk factors and complications (Table 2). Diabetic patients were nearly three times more likely to return to the operating room than those without diabetes (OR=2.78, CI=1.31-5.88, $p<0.01$). Postoperative infection was associated with a history of smoking (OR=4.20, CI=1.21-14.61, $p<0.05$). It appeared that patients with a history of smoking were more than four times more likely to experience postoperative infection. Experiencing any type of complication was associated with operating institution (OR=0.44, CI=0.25-0.78, $p<0.01$), BMI (OR=0.94, CI=0.89-0.99, $p<0.05$) and diabetes (OR=2.33, CI=1.18-4.60, $p<0.05$).

Operating institution was also significantly associated with experiencing a complication within 12 months (OR=0.36, CI=0.18-0.70, $p<0.01$). The institution with slightly higher complication rates appeared to have operated on a patient sample with higher comorbidity rates (Table 3).

Complication prediction with XGBM models

XGBM models coupled with the SMOTE dataset demonstrated strong predictive performance (Table 4). These models demonstrated higher performance (validation AUC: 0.86-0.97) compared to models trained and validated on the original dataset (validation AUC: 0.57-0.69). Models based on the SMOTE dataset predicted high numbers of true positives and true negatives. Models trained on the SMOTE training dataset and validated on the non-SMOTE holdout sample demonstrated performance that was not substantially superior to the models trained on the original dataset.

ROC curves were generated by running the trained models on the holdout validation datasets. The ROC curves and corresponding AUC associated with the four SMOTE XGBM models showed strong performance (Figure 2).

Plotting feature importance

Feature importance metrics were plotted for each of the SMOTE XGBM models (Figure 3). Age, BMI, procedure side, gender, a diagnosis of Parkinson's disease, institution and comorbidities appeared to be the most influential predictive features associated with complications. Feature importance appeared to vary slightly by model. When plotting complicated cases in the original dataset according to BMI and age, cases clustered at approximately age 70 and a BMI of 24 (Figure 4).

Carrying out predictions on hypothetical patient data

A set of hypothetical patients is shown to demonstrate the output of the XGBM predictive models (Table 5). Risk thresholds similar to those developed in cardiology risk stratification research were applied to facilitate interpretation of model output (low=<10%, moderate=10-15%, high=16-50%, very high= >50%) [69].

DISCUSSION

This study found multiple clinical predictors of complications in DBS surgery using supervised machine learning algorithms. Logistic regression showed that patient BMI, diabetes and operating institution were significantly associated with all complications grouped together. Diabetics were almost three times more likely to return to the operating room. A history of smoking was significantly associated with postoperative infection.

The XGBM supervised learning algorithm demonstrated strong predictive performance. The results of this study suggested that XGBMs, coupled with a SMOTE oversampling method, may be employed to successfully overcome the class imbalance problem and effectively predict complication outcomes in DBS surgery. This method may be used to estimate any individual patient's risk of complications. Plotting feature importance demonstrated that age, BMI, gender, procedure side, a diagnosis of Parkinson's disease, the operating institution and preoperative comorbidities were influential predictors of postoperative complications. The results of this study that suggested associations between preoperative risk factors and postoperative adverse outcomes are supported by previous research demonstrating that many of these same factors are significantly associated with complication outcomes in DBS surgery [31,70,71] and in other forms of neurosurgery [18,23,24,72,73].

Surgeons often perceive patterns in their clinical practice. Machine learning algorithms seem to approximate well the intuition of the surgeon. Postoperative complications are likely to arise as a result of complex interactions between many risk factors [74]. While logistic regression has been deployed in the past to predict surgical outcomes [18,24], other supervised learning algorithms, including XGBM, may be better suited to modeling these complex nonlinear relationships [44,48,75].

This study has demonstrated one potential approach to addressing the class imbalance problem, which is a major issue in surgical risk stratification [18,19,24,31]. The approach employed here, applying SMOTE oversampling in conjunction with the XGBM supervised learning algorithm, produced encouraging results.

Simple linear relationships between risks and outcomes are intuitive. Linear and logistic regression generate statistical weights associated with each predictor and can be represented with a linear equation. These approaches offer rapid interpretability and an impression of understandability. In contrast, advanced supervised machine learning algorithms are often more complex, inscrutable and opaque. Surgeons are likely to have a lower level of trust in, and therefore may demonstrate weaker adoption of, opaque machine learning algorithms as decision support tools. The XGBM performance statistics, feature importance plots and hypothetical cases generated help to address this issue by providing some insight into the mechanics of the XGBM models developed. More work on developing the "explainability" of these models is required.

A collection of hypothetical cases was presented to demonstrate the risk stratification outputs of the supervised learning models developed. There may be a tendency to attempt to identify patterns in the hypothetical patient data displayed and the corresponding risk evaluation output statistics. However, this tendency is fraught because the number of hypothetical cases displayed is small and the algorithms are able to model complex nonlinearities in the data, based on hundreds of training cases, which are likely to evade human judgement. Similar to previous research [18], these examples provide a random selection of cases and patient characteristics to offer clinicians a general sense of the predictive risk outputs of the models trained. They are not intended to offer a systematic demonstration of the complex relationships modeled by the trained algorithms.

These machine learning models have the potential to facilitate patient safety improvements [76]. They may be used to stimulate a deeper conversation about complications with a patient prior to surgery, more attention throughout the process from the surgeon and surgical team, closer patient follow-up and activation of other organizational patient support processes postoperatively. Models of this nature should form part of a broader comprehensive approach to clinical risk stratification and patient safety improvement. As an example from another domain of neurosurgery, the Seattle Spine Team has developed a systematic and standardised approach that incorporates multidisciplinary patient review conferences, specialized clinical teams, intraoperative monitoring protocols and multi-surgeon operating practices, in addition to the development of experimental decision support systems underpinned by machine learning methods [24,77–84].

The advantages of using machine learning methods to stratify risk in neurosurgery are numerous. Machine learning methods are more capable of capturing complex nonlinearities in very large datasets than traditional statistical techniques and can be deployed to production using cloud computing services for potential use by clinicians and patients globally [85,86]. These tools are well-suited to high-volume complex data processing, they facilitate access to information, they save time and they have the potential to augment the clinical functioning of the neurosurgeon. Incorporating machine learning tools into the neurosurgical workflow may assist in reducing the likelihood of clinical error and positively engaging the patient. Supervised machine learning models can provide accurate and individualized outcome predictions, which are likely to be beneficial as healthcare progresses toward a future that is more precise and value based. Prediction datapoints may feed into and influence perioperative care processes and decisions or intraoperative treatment by human and robotic systems. On the other hand, it may not be suitable to apply machine learning methods to datasets that are erroneous, exceedingly noisy, obsolete or biased. In these cases, it may be preferable to rely on the unassisted judgment of the expert surgeon and an experienced clinical team.

Limitations and future research

The performance metrics using SMOTE oversampling and extreme gradient boosting were strong. Such high performance of the XGBM algorithms suggests that some degree of overfitting may have occurred, despite built-in overfitting mitigation. This, however, is difficult to assess, particularly in the context of limited case data. Caution and appropriate clinical judgement should be exercised if deploying and using these models to make predictions on new patient data. Further validation on new data from other institutions and larger datasets would be beneficial.

While assessing the effects of the use of intraoperative CT on complication rates and patient outcomes was beyond the scope of this study, it may have been beneficial to control for its use in the analysis. Per patient labelling of this variable was not captured in the dataset and this is therefore a limitation of this study. Similarly, it would be beneficial for future predictive modeling studies to control for additional preoperative clinical variables in multivariate analyses. These may include pre- and post-operative functional status, anemia, operating time, the number of electrode passes, passage through the ventricle and a patient history of coronary artery disease or stroke.

A primary aim of this study was to develop models capable of stratifying patient complication risk in DBS surgery. To achieve this and to mitigate the limitations of the dataset, complication subcategories were amalgamated into a superordinate variable representing general clinical risk and adverse outcomes. This approach allowed the development of a set of useful and applicable models. However, it must be noted that these models are broad in their risk predictions and that to predict specific types of complications, which would enable the implementation of specific clinical risk mitigation tactics (e.g., augmented infection prevention or operating room preparation for a returning patient), larger datasets and more modeling work are required. The variables included in the superordinate complication outcome variable fall logically under the banner of adverse postoperative clinical events. While the specific outcomes that make up this variable may be considered diverse, amalgamating them remains clinically useful because together they broadly indicate high risk patients that may require additional critical clinical thought and discussion, resources and careful perioperative management.

Future research may deploy the methods applied here for the prediction of complications associated with other surgical procedures that are characterized by a similar class imbalance problem. Studies may also develop supervised learning models to predict positive functional outcomes and the degree of functional improvement associated with various neurosurgical procedures. Future work may focus on the development of clinical decision support systems to be applied in clinical practice and to deliver decision-support benefits directly to patients via application to patient consultations in the clinics [87].

Conclusion

Significant complication risk factors were detected and supervised machine learning algorithms effectively predicted adverse outcomes in DBS surgery. These supervised learning models can be used for the improvement of risk stratification, preoperative patient informed consent and clinical planning to make DBS surgery safer for patients. XGBMs and SMOTE appear to be useful tools for the prediction of complication outcomes and risk stratification in DBS surgical practice.

REFERENCES

1. Rumalla, K., Smith, K. A., Follett, K., Nazzaro, J. & Arnold, P. M. Rates, Causes, Risk Factors, and Outcomes of Readmission Following Deep Brain Stimulation for Movement Disorders: Analysis of the US Nationwide Readmissions Database. *Clin. Neurol. Neurosurg.* (2018).

2. Zhou, J. J., Chen, T., Farber, S. H., Shetter, A. G. & Ponce, F. A. Open-loop deep brain stimulation for the treatment of epilepsy: a systematic review of clinical outcomes over the past decade (2008–present). *Neurosurg. Focus* **45**, E5 (2018).

3. Matias, C. M., Frizon, L. A., Nagel, S. J., Lobel, D. A. & Machado, A. G. Deep brain stimulation outcomes in patients implanted under general anesthesia with frame-based stereotaxy and intraoperative MRI. *J. Neurosurg.* 1–7 (2018).

4. Kleiner□Fisman, G. *et al.* Subthalamic nucleus deep brain stimulation: summary and meta□ analysis of outcomes. *Mov. Disord. Off. J. Mov. Disord. Soc.* **21**, S290–S304 (2006).

5. Mueller, J. *et al.* Pallidal deep brain stimulation improves quality of life in segmental and generalized dystonia: results from a prospective, randomized sham□controlled trial. *Mov. Disord.* **23**, 131–134 (2008).

6. Flora, E. Della, Perera, C. L., Cameron, A. L. & Maddern, G. J. Deep brain stimulation for essential tremor: a systematic review. *Mov. Disord.* **25**, 1550–1559 (2010).

7. Schrock, L. E. *et al.* Tourette syndrome deep brain stimulation: a review and updated recommendations. *Mov. Disord.* **30**, 448–471 (2015).

8. Herzog, J. *et al.* Most effective stimulation site in subthalamic deep brain stimulation for Parkinson’s disease. *Mov. Disord.* **19**, 1050–1054 (2004).

9. Odekerken, V. J. J. *et al.* Subthalamic nucleus versus globus pallidus bilateral deep brain stimulation for advanced Parkinson’s disease (NSTAPS study): a randomised controlled trial. *Lancet Neurol.* **12**, 37–44 (2013).

10. Schlaepfer, T. E., Bewernick, B. H., Kayser, S., Mädler, B. & Coenen, V. A. Rapid effects of

deep brain stimulation for treatment-resistant major depression. *Biol. Psychiatry* **73**, 1204–1212 (2013).

11. Little, S. *et al.* Adaptive deep brain stimulation in advanced Parkinson disease. *Ann. Neurol.* **74**, 449–457 (2013).

12. Miocinovic, S., Somayajula, S., Chitnis, S. & Vitek, J. L. History, applications, and mechanisms of deep brain stimulation. *JAMA Neurol.* **70**, 163–171 (2013).

13. Figee, M. *et al.* Deep brain stimulation restores frontostriatal network activity in obsessive-compulsive disorder. *Nat. Neurosci.* **16**, 386–387 (2013).

14. Rodriguez-Oroz, M. C. *et al.* Bilateral deep brain stimulation in Parkinson's disease: a multicentre study with 4 years follow-up. *Brain* **128**, 2240–2249 (2005).

15. Kalakoti, P. *et al.* Predictors of unfavorable outcomes following deep brain stimulation for movement disorders and the effect of hospital case volume on outcomes: an analysis of 33, 642 patients across 234 US hospitals using the National (Nationwide) Inpatient Sample from 20. *Neurosurg. Focus* **38**, E4 (2015).

16. Zrinzo, L., Foltynie, T., Limousin, P. & Hariz, M. I. Reducing hemorrhagic complications in functional neurosurgery: a large case series and systematic literature review. *J. Neurosurg.* **116**, 84–94 (2012).

17. Hardaway, F. A., Raslan, A. M. & Burchiel, K. J. Deep brain stimulation-related infections: analysis of rates, timing, and seasonality. *Neurosurgery* **83**, 540–547 (2017).

18. Khor, S. *et al.* Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. *JAMA Surg.* **153**, 634–642 (2018).

19. Buchlak, Q. D. *et al.* Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. *Neurosurg. Rev.* 1–19 (2019). doi:https://doi.org/10.1007/s10143-019-01163-8

20. Senders, J. T. *et al.* Machine learning and neurosurgical outcome prediction: a systematic

review. *World Neurosurg.* **109**, 476–486 (2018).

21. Boachie-Adjei, O. *et al.* Surgical Risk Stratification Based on Preoperative Risk Factors in Severe Pediatric Spinal Deformity Surgery. *Spine Deform.* **2**, 340–349 (2014).

22. Bilimoria, K. Y. *et al.* Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J. Am. Coll. Surg.* **217**, 833-842. e3 (2013).

23. Bekelis, K., Desai, A., Bakhoum, S. F. & Missios, S. A predictive model of complications after spine surgery: The National Surgical Quality Improvement Program (NSQIP) 2005-2010. *Spine J.* **14**, 1247–1255 (2014).

24. Buchlak, Q. D. *et al.* The Seattle spine score: Predicting 30-day complication risk in adult spinal deformity surgery. *J. Clin. Neurosci.* (2017). doi:10.1016/j.jocn.2017.06.012

25. Stopa, B. M., Yan, S. C., Dasenbrock, H. H., Kim, D. H. & Gormley, W. B. Variance reduction in neurosurgical practice: The case for analytics driven decision support in the era of Big Data. *World Neurosurg.* (2019).

26. Angeles, P., Tai, Y., Pavese, N., Wilson, S. & Vaidyanathan, R. Automated assessment of symptom severity changes during deep brain stimulation (DBS) therapy for Parkinson's disease. in *2017 International Conference on Rehabilitation Robotics (ICORR)* 1512–1517 (IEEE, 2017).

27. Kostoglou, K. *et al.* Classification and Prediction of Clinical Improvement in Deep Brain Stimulation From Intraoperative Microelectrode Recordings. *IEEE Trans. Biomed. Eng.* **64**, 1123–1130 (2017).

28. Taghva, A. Hidden semi-Markov models in the computerized decoding of microelectrode recording data for deep brain stimulator placement. *World Neurosurg.* **75**, 758–763 (2011).

29. Taghva, A. An automated navigation system for deep brain stimulator placement using hidden Markov models. *Oper. Neurosurg.* **66**, ons-108 (2010).

30. Baumgarten, C. *et al.* Image-guided preoperative prediction of pyramidal tract side effect in deep brain stimulation: proof of concept and application to the pyramidal tract side effect induced by pallidal stimulation. *J. Med. Imaging* **3**, 25001 (2016).

31. Buchlak, Q. D., Kowalczyk, M., Leveque, J.-C., Wright, A. & Farrokhi, F. Risk stratification in deep brain stimulation surgery: Development of an algorithm to predict patient discharge disposition with 91.9% accuracy. *J. Clin. Neurosci.* (2018).

32. Wong, S., Baltuch, G. H., Jaggi, J. L. & Danish, S. F. Functional localization and visualization of the subthalamic nucleus from microelectrode recordings acquired during DBS surgery with unsupervised machine learning. *J. Neural Eng.* **6**, 26006 (2009).

33. Valsky, D. *et al.* Stop! border ahead: Automatic detection of subthalamic exit during deep brain stimulation surgery. *Mov. Disord.* **32**, 70–79 (2017).

34. Zaidel, A., Spivak, A., Shpigelman, L., Bergman, H. & Israel, Z. Delimiting subterritories of the human subthalamic nucleus by means of microelectrode recordings and a Hidden Markov Model. *Mov. Disord.* **24**, 1785–1793 (2009).

35. Zhang, Y. & Haghani, A. A gradient boosting method to improve travel time prediction. *Transp. Res. Part C Emerg. Technol.* **58**, 308–324 (2015).

36. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 1189–1232 (2001).

37. Lu, J. *et al.* Estimation of elimination half-lives of organic chemicals in humans using gradient boosting machine. *Biochim. Biophys. Acta (BBA)-General Subj.* **1860**, 2664–2671 (2016).

38. Touzani, S., Granderson, J. & Fernandes, S. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy Build.* **158**, 1533–1543 (2018).

39. Rawi, R. *et al.* PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics* **34**, 1092–1098 (2017).

40. Fan, J. *et al.* Comparison of Support Vector Machine and Extreme Gradient Boosting for

predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Convers. Manag.* **164**, 102–111 (2018).

41. Esmaili, N., Piccardi, M., Kruger, B. & Girosi, F. Analysis of healthcare service utilization after transport-related injuries by a mixture of hidden Markov models. *PLoS One* **13**, e0206274 (2018).

42. Brown, I. & Mues, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst. Appl.* **39**, 3446–3453 (2012).

43. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 1263–1284 (2008).

44. Raschka, S. & Mirjalili, V. *Python machine learning*. (Packt Publishing Ltd, 2017).

45. Grigsby, J., Kramer, R. E., Schneiders, J. L., Gates, J. R. & Brewster Smith, W. Predicting outcome of anterior temporal lobectomy using simulated neural networks. *Epilepsia* **39**, 61–66 (1998).

46. Abouzari, M., Rashidi, A., Zandi-Toghani, M., Behzadi, M. & Asadollahi, M. Chronic subdural hematoma outcome prediction using logistic regression and an artificial neural network. *Neurosurg. Rev.* (2009). doi:10.1007/s10143-009-0215-3

47. Patel, J. L. & Goyal, R. K. Applications of artificial neural networks in medical science. *Curr. Clin. Pharmacol.* **2**, 217–226 (2007).

48. Kim, J. S. *et al.* Predicting Surgical Complications in Patients Undergoing Elective Adult Spinal Deformity Procedures Using Machine Learning. *Spine Deform.* **6**, 762–770 (2018).

49. Japkowicz, N. & Stephen, S. The class imbalance problem: A systematic study. *Intell. data Anal.* **6**, 429–449 (2002).

50. Chawla, N. V. Data mining for imbalanced datasets: An overview. in *Data mining and knowledge discovery handbook* 875–886 (Springer, 2009).

51. Japkowicz, N. The class imbalance problem: Significance and strategies. in *Proc. of the Int'l*

Conf. on Artificial Intelligence (2000).

52. Mazurowski, M. A. *et al.* Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks* **21**, 427–436 (2008).

53. Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.* **30**, 25–36 (2006).

54. Sahin, Y., Bulkan, S. & Duman, E. A cost-sensitive decision tree approach for fraud detection. *Expert Syst. Appl.* **40**, 5916–5923 (2013).

55. Japkowicz, N., Myers, C. & Gluck, M. A novelty detection approach to classification. in *IJCAI* **1**, 518–523 (1995).

56. Kubat, M., Holte, R. C. & Matwin, S. Machine learning for the detection of oil spills in satellite radar images. *Mach. Learn.* **30**, 195–215 (1998).

57. Liu, X.-Y., Wu, J. & Zhou, Z.-H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man, Cybern. Part B* **39**, 539–550 (2008).

58. Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J. & Napolitano, A. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man, Cybern. A Syst. Humans* **40**, 185–197 (2009).

59. Chawla, N. V, Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

60. Carlson, J. D., McLeod, K. E., McLeod, P. S. & Mark, J. B. Stereotactic accuracy and surgical utility of the O-arm in deep brain stimulation surgery. *Oper. Neurosurg.* **13**, 96–107 (2016).

61. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* **338**, b2393 (2009).

62. Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. & Solenberger, P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv.*

Methodol. **27**, 85–96 (2001).

63. Horton, N. J. & Kleinman, K. P. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am. Stat.* **61**, 79–90 (2007).

64. Steyerberg, E. W. & van Veen, M. Imputation is beneficial for handling missing data in predictive models. *J. Clin. Epidemiol.* **60**, 979 (2007).

65. Batista, G. E. & Monard, M. C. An analysis of four missing data treatment methods for supervised learning. *Appl. Artif. Intell.* **17**, 519–533 (2003).

66. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 785–794 (ACM, 2016).

67. Senders, J. T. *et al.* Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery* (2017).

68. Saad, F. & Mansinghka, V. Probabilistic Data Analysis with Probabilistic Programming. 1–46 (2016).

69. Gabb, G. M. *et al.* Guideline for the diagnosis and management of hypertension in adults—2016. *Med. J. Aust.* **205**, 85–89 (2016).

70. Hu, K., Moses, Z. B., Hutter, M. M. & Williams, Z. Short-term adverse outcomes after deep brain stimulation treatment in patients with parkinson disease. *World Neurosurg.* **98**, 365–374 (2017).

71. Tanaka, M. *et al.* Risk Factors for Postoperative Delirium After Deep Brain Stimulation Surgery for Parkinson Disease. *World Neurosurg.* **114**, e518–e523 (2018).

72. Akins, P. T. *et al.* Risk factors associated with 30-day readmissions after instrumented spine surgery in 14,939 patients. *Spine (Phila. Pa. 1976).* **40**, 1022–1032 (2015).

73. Murphy, M. E. *et al.* Predictors of Discharge to a Nonhome Facility in Patients Undergoing Lumbar Decompression Without Fusion for Degenerative Spine Disease. *Neurosurgery* (2017).

74. Cooney, M. T., Dudina, A. L. & Graham, I. M. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. *J. Am. Coll. Cardiol.* **54**, 1209–1227 (2009).

75. Shi, H.-Y., Hwang, S.-L., Lee, K.-T. & Lin, C.-L. In-hospital mortality after traumatic brain injury surgery: a nationwide population-based comparison of mortality predictors used in artificial neural network and logistic regression models. *J. Neurosurg.* **118**, 746–752 (2013).

76. Bates, D. W. & Gawande, A. A. Improving safety with information technology. *N. Engl. J. Med.* **348**, 2526–2534 (2003).

77. Buchlak, Q. D., Yanamadala, V., Leveque, J.-C. & Sethi, R. Complication avoidance with pre-operative screening: insights from the Seattle spine team. *Curr. Rev. Musculoskelet. Med.* **9**, (2016).

78. Sethi, R. K. *et al.* The Seattle Spine Team approach to adult deformity surgery: a systems-based approach to perioperative care and subsequent reduction in perioperative complication rates. *Spine Deform.* **2**, 95–103 (2014).

79. Sethi, R. *et al.* A systematic multidisciplinary initiative for reducing the risk of complications in adult scoliosis surgery. *J. Neurosurg. Spine* **26**, (2017).

80. Sethi, R. K., Buchlak, Q. D., Leveque, J.-C., Wright, A. K. & Yanamadala, V. V. Quality and safety improvement initiatives in complex spine surgery. in *Seminars in Spine Surgery* **30**, 111–120 (Elsevier, 2018).

81. Yanamadala, V. *et al.* Multidisciplinary Evaluation Leads to the Decreased Utilization of Lumbar Spine Fusion. *Spine (Phila. Pa. 1976).* **42**, (2017).

82. Bauer, J. M., Yanamadala, V., Shah, S. A. & Sethi, R. K. Two Surgeon Approach for Complex Spine Surgery: Rationale, Outcome, Expectations, and the Case for Payment Reform. *JAAOS-Journal Am. Acad. Orthop. Surg.* **27**, e408–e413 (2019).

83. Buchlak, Q. D., Yanamadala, V., Leveque, J.-C. & Sethi, R. Preoperative Clinical Evaluation of

Adult Lumbar Scoliosis. in *Adult Lumbar Scoliosis* 61–70 (Springer, 2017).

84. Yanamadala, V., Buchlak, Q., Leveque, J. C. & Sethi, R. How to Decrease Complications in the Management of Adult Lumbar Scoliosis. *Instr. Course Lect.* **66**, 379–390 (2017).

85. Dalbhanjan, P. Overview of deployment options on AWS. *Amaz. Whitepapers* (2015).

86. Copeland, M., Soh, J., Puca, A., Manning, M. & Gollob, D. Microsoft azure and cloud computing. in *Microsoft Azure* 3–26 (Springer, 2015).

87. Kawamoto, K., Houlihan, C. A., Balas, E. A. & Lobach, D. F. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj* **330**, 765 (2005).

Journal Pre-proof

Figure legend:

- *Figure 1: Schematic outline of the two main phases of the analysis process.*
- *Figure 2: ROC curves for each of the SMOTE XGBM models, derived from the holdout test validation datasets.*
- *Figure 3: Feature importance plots for each of the SMOTE XGBM models. BMI = body mass index. DM diabetes mellitus. GPi = globus pallidus. HB = hemiballismus. HTN = hypertension. L = left. PD = Parkinson's Disease. R = right. STN = subthalamic nucleus. VIM = ventral intermedius nucleus.*
- *Figure 4: Joint plots of complicated cases (any complication; A) and uncomplicated cases (B) in our sample according to age and BMI. Complicated cases clustered at approximately age=70 and BMI=24, whereas uncomplicated cases clustered at approximately age=69 and BMI=28. Histograms plot age and BMI frequency distributions.*

Feature Category	Feature	Feature class	Count (%)
Predictors	Institution	Institution 1	201 (40%)
		Institution 2	300 (60%)
	Age	75 and over	70 (14%)
		Under 75	431 (86%)
	Gender	Male	318 (63%)
		Female	183 (37%)
	Diagnoses	Parkinson’s disease	349 (70%)
		Essential tremor	129 (26%)
		Dystonia	11 (2%)
		Other	12 (2%)
	BMI	≥25	335 (67%)
		18 to 24.9	157 (31%)
		<18	9 (2%)
	Comorbidities and risk factors	Smoking history	25 (5%)
		Immune suppressed	25 (5%)
		Diabetes	67 (13%)
		Hypertension	231 (46%)
	Procedure type	Subthalamic (STN)	349 (70%)
		Thalamic (VIM)	128 (26%)
		Globus pallidus internus (GPi)	22 (4%)
		Other	2 (0%)
Outcomes	Intracranial hemorrhage		15 (3%)
	Readmission		17 (3%)
	Ischemic infarction		3 (1%)
	Seizure		3 (1%)
	Lead fractures		18 (4%)
	Electrode migration		8 (2%)
	Battery loose or flipping		7 (1%)
	Device malfunction		26 (5%)
	Return to operating room		53 (11%)
	Infection		27 (5%)
	Hemiparesis		5 (1%)
	Facial droop		6 (1%)
	Sensory change		4 (1%)
	Complication other		8 (2%)
	Complication any		83 (17%)
	Complication within 12 months		59 (12%)

Table 1: Descriptive statistics displaying the classes of each of the predictors and outcome features in the dataset of 501 DBS patients. Other diagnoses included cluster headache, Holmes tremor and Tourette Syndrome.

		Any complication		Complication within 12 months		Return to the operating room		Infection	
		Coefficient	OR (95% CI)	Coefficient	OR (95% CI)	Coefficient	OR (95% CI)	Coefficient	OR (95% CI)
Intercept		0.35	1.55 (0.35, 6.90)	-0.60	0.55 (0.10, 3.00)	-0.79	0.46 (0.08, 2.67)	-2.20	0.11 (0.01, 1.11)
Demographics	Institution 02	**-0.82****	**0.44 (0.25, 0.78)**	**-1.03****	**0.36 (0.18, 0.70)**	-0.39	0.68 (0.35, 1.34)	--	--
	Age 75 and over	0.44	1.55 (0.77, 3.13)	0.53	1.70 (0.75, 3.84)	0.17	1.18 (0.50, 2.80)	0.90	2.45 (0.88, 6.78)
	Male	-0.09	0.91 (0.55, 1.51)	0.06	1.06 (0.58, 1.91)	-0.09	0.91 (0.50, 1.68)	0.13	1.14 (0.48, 2.68)
	BMI at implant	**-0.07***	**0.94 (0.89, 0.99)**	-0.05	0.95 (0.90, 1.01)	-0.04	0.96 (0.90, 1.02)	-0.06	0.95 (0.87, 1.03)
Clinical features	Diabetes	**0.84***	**2.33 (1.18, 4.60)**	0.78	2.17 (0.98, 4.80)	**1.02****	**2.78 (1.31, 5.88)**	0.56	1.75 (0.58, 5.29)
	Hypertension	0.00	1.00 (0.58, 1.73)	0.21	1.23 (0.65, 2.32)	-0.18	0.84 (0.43, 1.60)	0.83	2.29 (0.99, 5.30)
	Smoking history	0.16	1.18 (0.40, 3.46)	0.38	1.46 (0.45, 4.79)	0.27	1.31 (0.41, 4.25)	**1.44***	**4.20 (1.21, 14.61)**
	Immunosuppression	0.14	1.15 (0.38, 3.55)	0.35	1.42 (0.43, 4.67)	-1.30	0.27 (0.03, 2.27)	--	--
	ET	0.02	1.02 (0.23, 4.55)	0.53	1.70 (0.43, 6.67)	-0.02	0.98 (0.19, 5.09)	-0.45	0.64 (0.07, 5.96)
	Dystonia	-1.02	0.36 (0.03, 4.12)	--	--	-0.53	0.59 (0.05, 7.25)	--	--
	Diagnosis other	-0.59	0.55 (0.07, 4.17)	--	--	--	--	--	--
	Thalamic (Vim)	-0.10	0.91 (0.21, 4.04)	-1.11	0.33 (0.08, 1.33)	0.37	1.44 (0.28, 7.55)	-0.47	0.62 (0.07, 5.84)
	Globus pallidus internus (GPi)	0.58	1.78 (0.46, 6.97)	0.20	1.22 (0.32, 4.70)	0.60	1.82 (0.39, 8.51)	0.10	1.10 (0.21, 5.70)
	Left sided procedure	0.20	1.23 (0.65, 2.32)	0.50	1.65 (0.80, 3.38)	-0.05	0.95 (0.44, 2.07)	0.25	1.29 (0.46, 3.62)
	Right sided procedure	-0.23	0.79 (0.34, 1.85)	0.19	1.20 (0.49, 2.98)	-0.52	0.59 (0.20, 1.75)	0.32	1.37 (0.41, 4.62)
LLR p-value			p=0.09		p<0.05		p=0.54		p=0.21

*Table 2: Multivariate logistic regression modelling results. These results are based on analysis of the original (non-SMOTE) dataset. CI = confidence interval. LLR = log likelihood ratio. OR = odds ratio. The reference categories were: female, Parkinson's Disease, age <75, institution 01, an operation conducted on both sides and an STN procedure. *p<0.05, **p<0.01.*

		Institution 01	Institution 02
Number of cases		201	300
Demographics	Age (mean, SD)	62.12 (10.52)	66.28 (9.94)
	BMI (mean, SD)	27.71 (5.47)	27.17 (4.62)
	Female	41%	34%
Diagnosis	Parkinson’s disease	68%	71%
	Essential tremor	24%	27%
	Dystonia	1%	3%
Clinical features	Smoking history	5%	5%
	Immune suppressed	9%	2%
	Diabetes mellitus	13%	13%
	Hypertension	70%	30%
Target	STN	69%	70%
	VIM	27%	24%
	GPi	3%	5%
Procedure side	Left	36%	25%
	Right	17%	8%
	Both	47%	67%
Complication outcomes	Any complication	22%	13%
	Complication at 12 months	18%	7%
	Return to the operating room	11%	10%
	Infection	6%	5%

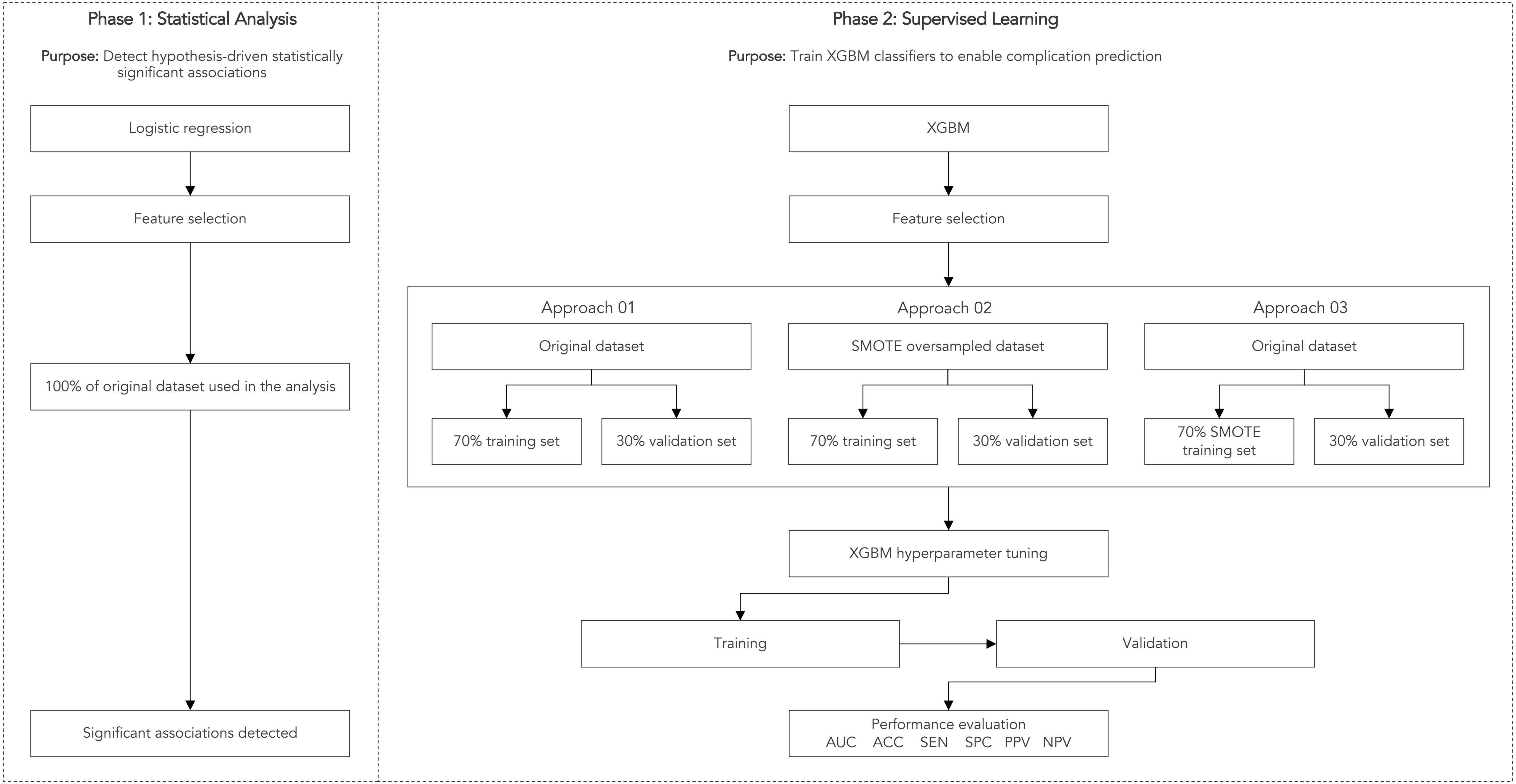
Table 3: A comparison of case characteristics between institutions, demonstrating a notable difference in the prevalence of patients with hypertension and immune suppression.

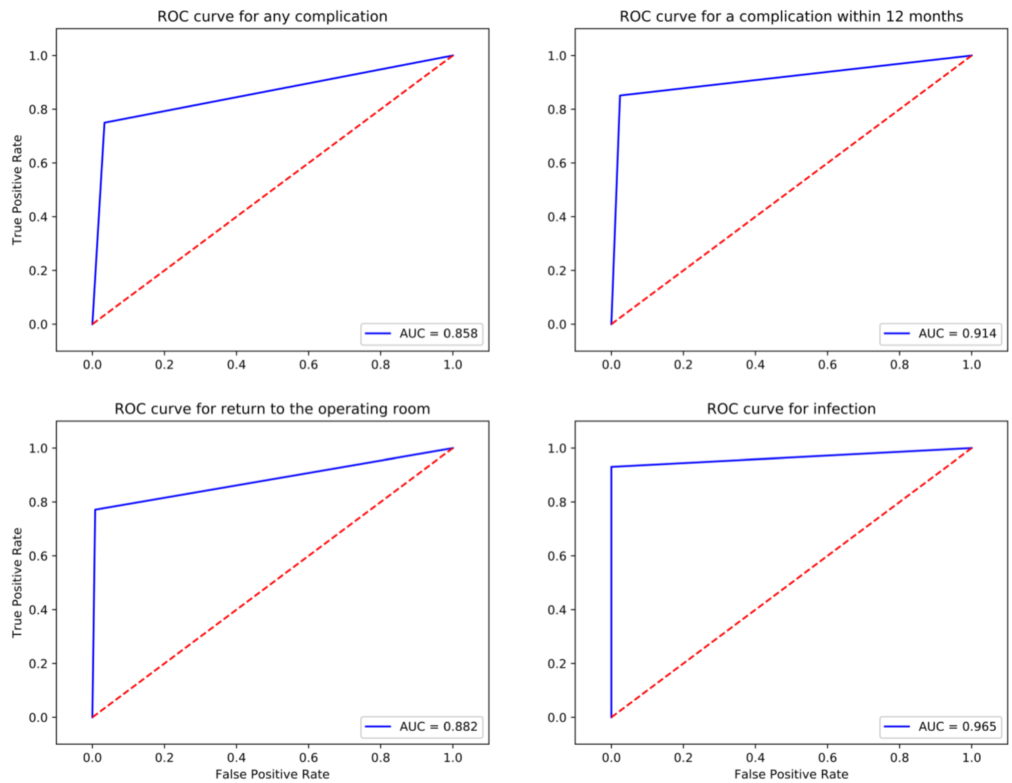
Data	Original Dataset								SMOTE Dataset							
Model	Any complication		Complication at 12 months		Return to operating room		Infection		Any complication		Complication at 12 months		Return to operating room		Infection	
Performance on validation holdout datasets																
Accuracy	0.66		0.86		0.88		0.95		0.85		0.91		0.88		0.97	
AUC	0.58		0.69		0.57		0.68		0.94		0.96		0.97		0.99	
Sensitivity	0.07		0.00		0.00		0.00		0.96		0.98		0.99		1.00	
Specificity	0.81		0.88		0.91		0.95		0.78		0.85		0.80		0.93	
PPV	0.08		0.00		0.00		0.00		0.75		0.85		0.77		0.93	
Confusion matrices																
Predicted	Actual								Actual							
	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
+	2	23	0	18	0	13	0	8	99	33	120	21	108	32	134	10
-	29	97	3	130	5	133	0	143	4	115	3	122	1	128	0	141

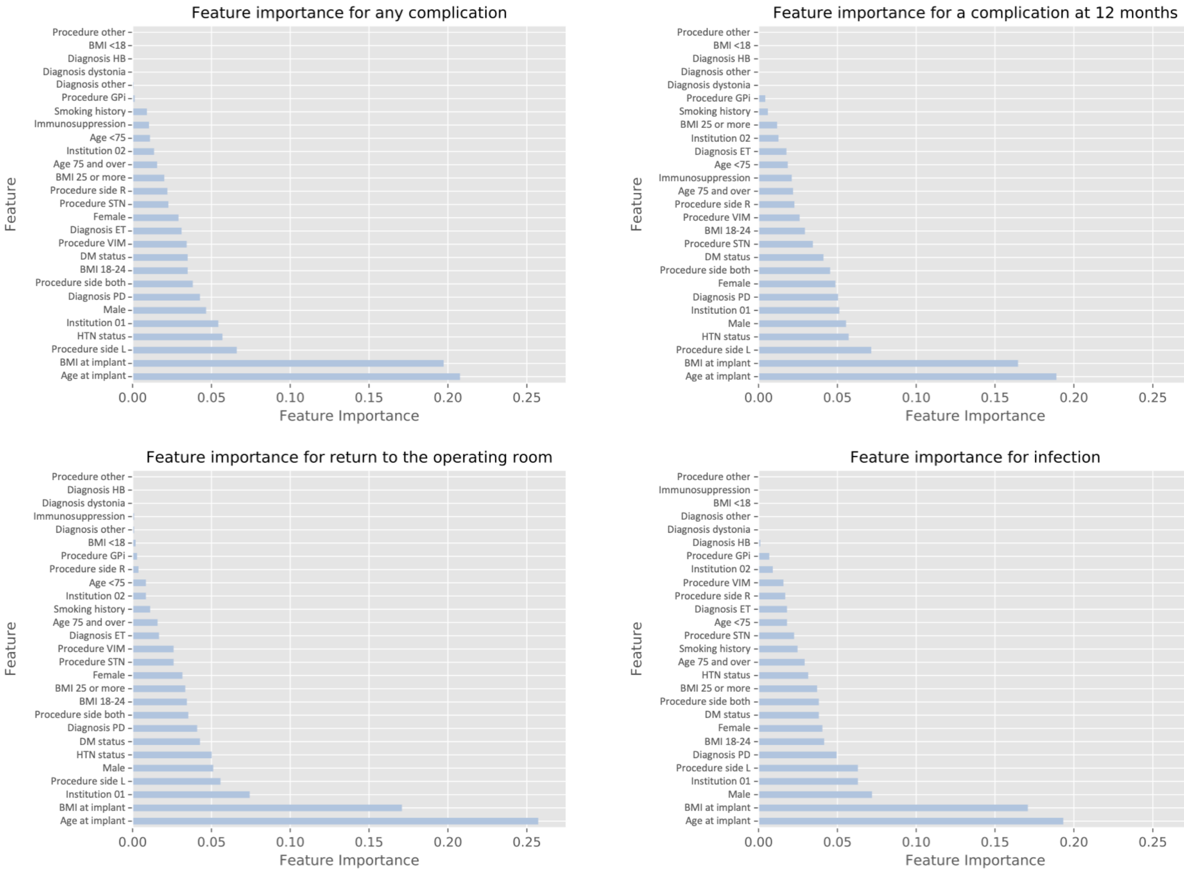
Table 4: Predictive performance metrics of XGBM models predicting (1) any complication, (2) complication within 12 months, (3) return to the operating room and (4) infection; using (a) the original dataset, and (b) the SMOTE oversampled training and validation datasets.

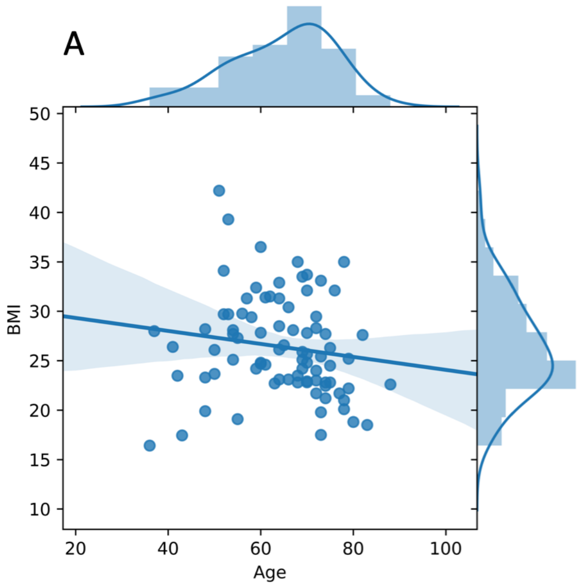
	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10
Institution	1	1	1	1	2	2	2	2	1	2
Age at implant	33	48	63	75	33	48	54	78	57	76
Gender	F	F	M	M	M	M	F	F	M	M
BMI	18	24	30	27	28	35	22	29	35	40
Diagnosis	PD	Dyst	PD	PD	PD	ET	PD	PD	PD	ET
Smoking history	No	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No
Immunosuppression	Normal	Normal	Normal	Yes	Normal	Normal	Yes	Normal	Yes	Normal
Diabetes status	--	DM	--	--	--	DM	--	DM	DM	DM
Hypertension status	--	--	HTN	--	HTN	--	--	HTN	HTN	--
Procedure target	STN	GPi	GPi	STN	STN	VIM	STN	STN	STN	VIM
Procedure side	Left	Both	Left	Right	Both	Both	Left	Both	Both	Right
Predicted probabilities of complication outcomes (likelihood shown in parentheses)										
Infection	M (11%)	H (25%)	L (7%)	H (29%)	H (17%)	L (4%)	M (10%)	M (12%)	H (16%)	H (19%)
Return to the operating room	M (12%)	VH (64%)	L (7%)	L (5%)	H (28%)	M (13%)	M (13%)	M (11%)	H (22%)	L (7%)
Any postoperative complication	H (43%)	VH (61%)	L (2%)	M (13%)	H (39%)	M (10%)	M (13%)	M (10%)	H (22%)	H (16%)
Postoperative complication within 12 months	M (14%)	H (19%)	L (2%)	H (32%)	L (6%)	L (8%)	H (17%)	L (3%)	H (23%)	H (31%)

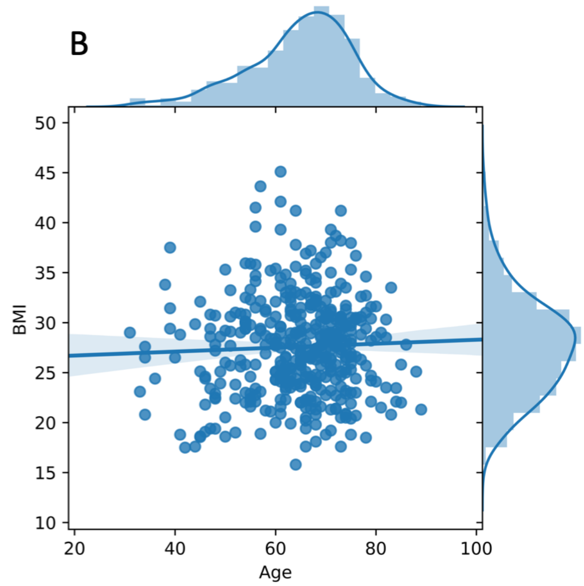
Table 5: Hypothetical patient characteristics and corresponding predicted complication likelihood. Risk thresholds are based on decision boundaries developed in cardiology: Low = <10%; Moderate = 10-15%; High = 16-50%; Very high = >50%. Dyst = dystonia. ET = essential tremor. GPi = globus pallidus. H = high. L = low. M = moderate. PD = Parkinson's Disease. STN = subthalamic nucleus. VH = very high. VIM = ventral intermedius nucleus.


Phase 1: Statistical Analysis
Purpose: Detect hypothesis-driven statistically significant associations
Logistic regression
Feature selection
100% of original dataset used in the analysis
Significant associations detected
Phase 2: Supervised Learning
Purpose: Train XGBM classifiers to enable complication prediction
XGBM
Feature selection
Approach 01
Original dataset
70% training set
30% validation set
Approach 02
SMOTE oversampled dataset
70% training set
30% validation set
Approach 03
Original dataset
70% SMOTE training set
30% validation set
XGBM hyperparameter tuning
Training
Validation
Performance evaluation
AUC ACC SEN SPC PPV NPV


ROC curve for any complication
ROC curve for a complication within 12 months
ROC curve for return to the operating room
ROC curve for infection
True Positive Rate
False Positive Rate
AUC = 0.858
AUC = 0.914
AUC = 0.882
AUC = 0.965
0.0
0.2
0.4
0.6
0.8
1.0


Feature importance for any complication
Procedure other
BMI <18
Diagnosis HB
Diagnosis dystonia
Diagnosis other
Procedure GPi
Smoking history
Immunosuppression
Age <75
Institution 02
Age 75 and over
BMI 25 or more
Procedure side R
Procedure STN
Female
Diagnosis ET
Procedure VIM
DM status
BMI 18-24
Procedure side both
Diagnosis PD
Male
Institution 01
HTN status
Procedure side L
BMI at implant
Age at implant
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Feature Importance
Feature importance for a complication at 12 months
Procedure other
BMI <18
Diagnosis HB
Diagnosis other
Diagnosis dystonia
Procedure GPi
Smoking history
BMI 25 or more
Institution 02
Diagnosis ET
Age <75
Immunosuppression
Age 75 and over
Procedure side R
Procedure VIM
BMI 18-24
Procedure STN
DM status
Procedure side both
Female
Diagnosis PD
Institution 01
Male
HTN status
Procedure side L
BMI at implant
Age at implant
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Feature Importance
Feature importance for return to the operating room
Procedure other
Diagnosis HB
Diagnosis dystonia
Immunosuppression
Diagnosis other
BMI <18
Procedure GPi
Procedure side R
Age <75
Institution 02
Smoking history
Age 75 and over
Diagnosis ET
Procedure VIM
Procedure STN
Female
BMI 25 or more
BMI 18-24
Procedure side both
Diagnosis PD
DM status
HTN status
Male
Procedure side L
Institution 01
BMI at implant
Age at implant
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Feature Importance
Feature importance for infection
Procedure other
Immunosuppression
BMI <18
Diagnosis other
Diagnosis dystonia
Diagnosis HB
Procedure GPi
Institution 02
Procedure VIM
Procedure side R
Diagnosis ET
Age <75
Procedure STN
Smoking history
Age 75 and over
HTN status
BMI 25 or more
Procedure side both
DM status
Female
BMI 18-24
Diagnosis PD
Procedure side L
Institution 01
Male
BMI at implant
Age at implant
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Feature Importance


A
BMI
Age
50
45
40
35
30
25
20
15
10
20
40
60
80
100


B
BMI
Age
50
45
40
35
30
25
20
15
10
20
40
60
80
100

Abbreviations list

Abbreviation	Expansion / Meaning
AUC	area under the receiver operating characteristics curve
CI	confidence interval
DBS	deep brain stimulation
Dyst	dystonia
ET	essential tremor
GBM	gradient boosting machine
GPi	globus pallidus
H	High
L	Low
LLR	log likelihood ratio
M	Moderate
NPV	negative predictive value
OR	odds ratio
PD	Parkinson's disease
PPV	positive predictive value
ROC	receiver operating characteristics
SMOTE	Synthetic Minority Oversampling Technique
STN	subthalamic nucleus
VH	Very high
VIM	ventral intermedius nucleus
XGBM	extreme gradient boosting machine