Delirium Prediction using Machine Learning Models on Preoperative Electronic Health Records Data

Anis Davoudi,
Department of Biomedical Engineering, University of Florida, Gainesville, USA

Ashkan Ebadi,
Department of Biomedical Engineering, University of Florida, Gainesville, USA

Parisa Rashidi,
Department of Biomedical Engineering, University of Florida, Gainesville, USA

Tazcan Ozrazgat-Baslanti,
Department of Medicine, University of Florida, Gainesville, USA

Azra Bihorac, and
Department of Medicine, University of Florida, Gainesville, USA

Alberto C. Bursian
Department of Anesthesiology, University of Florida, Gainesville, USA

Abstract

Electronic Health Records (EHR) are mainly designed to record relevant patient information during their stay in the hospital for administrative purposes. They additionally provide an efficient and inexpensive source of data for medical research, such as patient outcome prediction. In this study, we used preoperative Electronic Health Records to predict postoperative delirium. We compared the performance of seven machine learning models on delirium prediction: linear models, generalized additive models, random forests, support vector machine, neural networks, and extreme gradient boosting. Among the models evaluated in this study, random forests and generalized additive model outperformed the other models in terms of the overall performance metrics for prediction of delirium, particularly with respect to sensitivity. We found that age, alcohol or drug abuse, socioeconomic status, underlying medical issue, severity of medical problem, and attending surgeon can affect the risk of delirium.

Keywords

Delirium; electronic health record; machine learning; prediction

anisdavoudi@ufl.edu, ashkan.ebadi@gmail.com, parisa.rashidi@bme.ufl.edu, tezcan@phhp.ufl.edu, abihorac@ufl.edu, abursian@anest.ufl.edu.

I. Introduction

Delirium is a common transient neuropsychiatric disorder exhibited abruptly with fluctuations in consciousness and mental status (1). Delirium is linked to multiple adverse events, including increased morbidity and mortality, prolonged stays in the intensive care unit (ICU), and prolonged mechanical ventilation (2).

The prevalence of delirium is as high as 73% among surgical ICU patients, costing US between $38 and $150 billion per year (3, 4). Studies show that one third of delirium cases can benefit from multifactorial preventive measures and treatments (5–9). The high prevalence rate and the potential for successful intervention calls for accura te prediction methods to identify patients at higher risk of developing delirium. Previous studies show that delirium is currently under-recognized (10), thus preventing timely detection and treatment measures. Risk prediction for postoperative delirium is limited to physician's subjective risk assessment models that often rely on elaborate data extraction (11). To remedy this problem, machine learning models can be used to predict risk of delirium. The pathogenesis of postoperative delirium is not completely determined yet. However, there are many factors that are shown to contribute to higher risk of delirium development. These factors include, but are not limited to: age, admission type, primary surgical procedure. These factors are recorded in electronic health records and have been used in the literature for delirium prediction (6, 12, 13).

In this study, we will use preoperative Electronic Health Record (EHR) data for delirium prediction. EHR data are routinely collected for all patients at admission, and many studies have used them for prediction of health outcomes during and after the hospital stay (14). Previous works that have used EHR data for delirium prediction, have mainly used multivariate regression models due to their ease of interpretation and analysis (12, 13). Using logistic regression analysis, several other studies have identified important features in the preoperative dataset contributing to development of delirium, with up to 87% area under the receiver operating characteristics (ROC) curve (AUC) (12, 13, 15). Although promising, they mostly used additional features such as Mini-Mental State Examination score, Visual acuity, and Geriatric Depression Scale, which are not routine assessments in the hospitals. Others have been applied to specific populations such as older patients, elderly patients with hip surgery, or patients in the ICU (12, 13, 16).

In recent years, machine learning techniques have been used increasingly in medical research for analyzing complex medical data (17–21). In this paper, we have used the preoperative data routinely collected at admission for developing machine learning models to predict delirium. Unlike previous work, we do not rely on special assessments in our model. Thus, our model can be readily used in practice, using only EHR data. Our results also show better predictive performance compared to classification models in the literature. We used several methods including downsampling and Synthetic Minority Over-Sampling Technique (SMOTE) (22) to counter the imbalance in the outcome and improve the performance of the models. The rest of the paper is as follows: we explain the methods used in section II and present the results in section III, while section IV includes the discussion and conclusions.

II.   methods

A.   Participants

This study was approved by the University of Florida Institutional Review Board and Privacy Office as an exempt study with waiver of informed consent. We included all patients 18 years of age and older who were admitted for longer than 24 hours following any type of inpatient operative procedure between January 1, 2000 and November 30, 2010. The final cohort consisted of 51,457 patients. The main outcome of this study was delirium occurring at any time during the patient's stay in the hospital. Delirium was identified using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes (23).

Based on literature review and clinical expertise, we derived a set of seventy preoperative predictor features out of 285 available preoperative demographics, socio-economic, administrative, clinical, pharmacy, and laboratory variables. Patient comorbidities were derived using up to fifty preoperative ICD-9-CM codes. We used validated methods to define binary comorbidity variables (24, 25), and Charlson comorbidity index as a composite measure for medical comorbidities (26). We extracted medications dispensed on the first day of admission using the RxNorms coding, and grouped them into drug classes according to the United States Department of Veterans Affairs National Drug File-Reference Terminology (NDF-RT) (27).

B.   Analysis

The predictive analytics workflow of the study is outlined in Figure 1. We performed several preprocessing steps on the dataset to improve the computational efficiency and robustness of prediction models. Data preprocessing included data cleaning with removal of outliers, imputation of missing data, and optimization of categorical and nominal variables (28). To address the risk of overfitting, we randomly split the data. In each run, 80% of the data was used for model development and 20% for testing. Prevalence of delirium was similar in each partition using the sampling design. We further split the development data into 80% for model training, and 20% for validation to tune parameters. We repeated the process 50 times to report performance measures and confidence intervals. In each run, data was reshuffled before splitting.

We trained the algorithm on the development cohorts, while the reported results were obtained from the test cohorts. We used 20% of the cohort as the test cohort (n= 10,291) in each of the 50 repeated 5-fold cross validation runs (resulted in 250 different cohorts). The overall sample size allows a maximum width of 95% confidence interval for area under the curve of 0.06 for each model when prevalence of predicted complication is 3%.

We compared seven predictive modeling approaches: Naïve Bayes (NB), generalized additive model (GAM), logistic regression (LR), support vector machine (SVM), random forests (RF), extreme gradient boosting (XGB), and neural networks (NN). Naive Bayes is commonly used as a simple generative model, a category of predictive models that learns the distribution of the input data using the Bayes rule. On the other hand, discriminative models learn a direct map from the input data to the response labels such as logistic regression and

generalized additive model (29). Logistic regression is widely used in medical literature, and has been used previously for prediction models for delirium (30, 31); with the predicted risk monotonically increasing or decreasing. Generalized additive models are additive regression models that can relax the monotonicity assumption of logistic models and offer the advantage of estimating non-linear risk functions for continuous variables.

Support vector machine, random forests, extreme gradient boosting, and neural networks are among widely used machine learning techniques, but they have not been used for prediction of delirium before. SVM performs classification by finding a separating decision boundary in the input feature space. Random forests constructs many decision trees and typically classifies data according to the mode of the decision trees. These decision trees are trained by splitting the dataset into subsets on a value at a node, repeating this process on each subset in a recursive manner. Random forests improve their performance by averaging over multiple decision trees trained on different parts of the dataset, and thus reducing the risk of overfitting. Extreme gradient boosting works as an ensemble of weaker prediction models (decision trees here) in an iterative fashion. At each iteration, a new model is built that adds an estimator to provide a better approximation than the previous iteration, and each model learns to correct the previous stage model.

Neural networks are machine learning models inspired by networks of biological neurons. They contain layers of simple computing nodes that operate as nonlinear summing algorithm, interconnected by weighted connection lines, with weights being adjusted with new training samples (32, 33).

We used the event probabilities calculated by the predictive models to classify patients into event and nonevent categories. We applied a cutoff defined as the value where the Youden's J Statistic is maximized (34). We used this cutoff from the training dataset on the test dataset. Because of imbalance between the two outcome classes, accuracy alone does not give a complete view of the models' discrimination performance. For our study, we compared the models using accuracy, AUC, sensitivity as the proportion of true positives over total positive targets, and specificity as the proportion of true negatives over total negative targets.

We calculated the importance of the based on mean decrease in Gini impurity index. Gini impurity index can be calculated as in Equation (1). Every time a split is made on a variable, Gini indices for the two descendent nodes are less than the Gini index of the parent node. The decrease in Gini index for each variable is calculated by adding up the decreases in Gini index for the variables over all trees in the forest.

$$G = \sum_{i=1}^{n_c} p_i (1 - p_i) \quad (1)$$

Here, n_c is the number of classes in the outcome and p_i is the ratio of the class.

We also used downsampling and SMOTE algorithm to remedy the imbalance in the dataset. Downsampling works by randomly selecting equal numbers of the minority class as the

majority class for training, thus creating a more balanced dataset for training in each fold. One issue with downsampling is that it could lead to loss of potentially important information (35). On the other hand, SMOTE algorithm multiplies the vector between K neighboring samples of a sample by a random value between 0 and 1, and adds the result to that sample, creating synthetic samples.

III. Results

Among 51,457 adult patients who underwent major inpatient surgery in a tertiary care academic center, 3.12% had delirium occurring at any time during their hospital stay (Table I). Several features were identified as important for delirium prediction (top 15 shown in Figure 2.). Importance of features patients' zip code, county, and primary insurance suggests the effect of socioeconomic status of the patients on risk of delirium. Number of medication on admission day, number of diagnoses on admission day, time of surgery from admission, and admitting service type hint at the effect of severity of the patient's health problems. Primary surgical procedure and major diagnosis category show the importance of underlying medical issue on risk of delirium. Attending surgeon feature can show the experience of the surgeon. Prevalence of delirium varied among different surgery types, likely reflecting effect of the underlying primary disease that may predispose delirium (Table II). Distribution of outcome and variables did not differ between developing and testing cohorts.

The predictive performance of the models is given in Table III. Generalized additive model, logistic regression, random forests, and extreme gradient boosting had better performance compared to Naïve Bayes model, neural network, and support vector machine models, all with AUCs above 0.77, except for SVM. Random forests and generalized additive model were the only models that had high performance in AUC, accuracy, sensitivity, and specificity.

IV. Discussion

In our study, we compared the performance of prediction models for delirium using EHR data. We studied seven models: logistic regression, generalized additive models, support vector machines, naïve Bayes, random forests, extreme gradient boosting, and neural networks. Among these models, random forests and generalized additive models were the best performing delirium prediction models. The data that we used for these models are available at the point of access to preoperative care and do not require specialized assessments or self-report information. We included complex variables, such as residency ZIP codes and attending doctor. ZIP codes can act as a surrogate of neighborhood socioeconomic characteristics, which has been shown to be associated with multiple disease and health behaviors (36–38). The performance of attending doctors can potentially be a factor of postoperative outcomes as well (39, 40).

For this study, we first applied a data cleaning step (28) to reduce errors by removing the outliers. The preprocessing step was carried out in consultation with clinicians who provided a good understanding of the nature of data and clinical needs, and made the algorithms more robust and efficient. Generalized additive model and random forests were the preferred

models in this study due to their accuracy, relative efficiency, and ability to account for non-linearity of variables. Generalized additive model is a data driven model that has the flexibility to capture non-monotonicity in the predicted outcome. Random forests, as a decision tree-based approach, is capable of capturing conditionality, the relations between the features, and nonlinear relationships between the features and the outcome. Since parameter tuning is critical for support vector machines, neural network, random forests, and extreme gradient boosting, further fine tuning of their parameters may potentially improve the current results.

We optimized our parameters and model based on maximum Youden's index rather than accuracy. As a result, downsampling and SMOMTE did not significantly improve the performance of random forests and generalized additive models. We used external validation to report the true performance of the models on unseen data. We also used random forests model to rank the features used in the model based on their mean reduction in Gini impurity index. Many of the features chosen by the model show the previously overlooked socioeconomic features and surgeon experience for delirium prediction.

These models could be applied at point of access to preoperative care. They do not rely on self-reported data and specialized testing, and were derived from whole population data, routinely collected in preoperative period. They can help healthcare staff in identifying patients at higher risk of such complications. These prediction models can also help patients make informed decisions about their surgical procedures and the risks involved. Prospective validation of the model on different populations can improve the model for implementation in real-time clinical workflow for automated and simplified risk stratification in the preoperative period. Future work also includes applying other methods of working with unbalanced dataset and using other classification methods. We also intend to study the performance of prediction models by adding intra-operative and post-operative features to capture any complication and the trajectory of the patient during their stay in the hospital.

REFERENCES

1. Miyagawa Y, Yokoyama Y, Fukuzawa S, Fukata S, Ando M, Kawamura T, et al. Risk Factors for Postoperative Delirium in Abdominal Surgery: A Proposal of a Postoperative Delirium Risk Score in Abdominal Surgery. Dig Surg 2017;34(2):95–102. Epub 2016/09/01. doi: 10.1159/000449044. PubMed PMID: 27576903. [PubMed: 27576903]

2. Fong TG, Tulebaev SR, Inouye SK. Delirium in elderly adults: diagnosis, prevention and treatment. Nat Rev Neurol 2009;5(4):210–20. Epub 2009/04/07. doi: 10.1038/nrneurol.2009.24. PubMed PMID: 19347026; PubMed Central PMCID: PMCPMC3065676. [PubMed: 19347026]

3. Pandharipande P, Cotton BA, Shintani A, Thompson J, Pun BT, Morris JA, Jr., et al. Prevalence and risk factors for development of delirium in surgical and trauma intensive care unit patients. J Trauma 2008;65(1):34–41. Epub 2008/06/27. doi: 10.1097/TA.0b013e31814b2c4d. PubMed PMID: 18580517; PubMed Central PMCID: PMCPMC3773485. [PubMed: 18580517]

4. Leslie DL, Marcantonio ER, Zhang Y, Leo-Summers L, Inouye SK. One-year health care costs associated with delirium in the elderly population. Arch Intern Med 2008;168(1):27–32. Epub 2008/01/16. doi: 10.1001/archinternmed.2007.4. PubMed PMID: 18195192; PubMed Central PMCID: PMCPMC4559525. [PubMed: 18195192]

5. Gusmao-Flores D, Salluh JI, Chalhub RA, Quarantini LC. The confusion assessment method for the intensive care unit (CAM-ICU) and intensive care delirium screening checklist (ICDSC) for the diagnosis of delirium: a systematic review and meta-analysis of clinical studies. Crit Care

2012;16(4):R115 Epub 2012/07/05. doi: 10.1186/cc11407. PubMed PMID: 22759376; PubMed Central PMCID: PMCPMC3580690. [PubMed: 22759376]

6. Ouimet S, Kavanagh BP, Gottfried SB, Skrobik Y. Incidence, risk factors and consequences of ICU delirium. Intensive Care Med 2007;33(1):66–73. Epub 2006/11/15. doi: 10.1007/s00134-006-0399-8. PubMed PMID: 17102966. [PubMed: 17102966]

7. Boettger S, Nunez DG, Meyer R, Richter A, Fernandez SF, Rudiger A, et al. Delirium in the intensive care setting: A reevaluation of the validity of the CAM-ICU and ICDSC versus the DSM-IV-TR in determining a diagnosis of delirium as part of the daily clinical routine. Palliat Support Care 2017:1–9. Epub 2017/02/09. doi: 10.1017/s1478951516001176. PubMed PMID: 28173895.

8. Davis DH, Kreisel SH, Muniz Terrera G, Hall AJ, Morandi A, Boustani M, et al. The epidemiology of delirium: challenges and opportunities for population studies. Am J Geriatr Psychiatry 2013;21(12):1173–89. Epub 2013/08/03. doi: 10.1016/j.jagp.2013.04.007. PubMed PMID: 23907068; PubMed Central PMCID: PMCPMC3837358. [PubMed: 23907068]

9. Milisen K, Steeman E, Foreman MD. Early detection and prevention of delirium in older patients with cancer. Eur J Cancer Care (Engl) 2004;13(5):494–500. Epub 2004/12/21. doi: 10.1111/j.1365-2354.2004.00545.x. PubMed PMID: 15606717. [PubMed: 15606717]

10. Ely EW, Shintani A, Truman B, Speroff T, Gordon SM, Harrell FE, Jr., et al. Delirium as a predictor of mortality in mechanically ventilated patients in the intensive care unit. Jama 2004;291(14):1753–62. Epub 2004/04/15. doi: 10.1001/jama.291.14.1753. PubMed PMID: 15082703. [PubMed: 15082703]

11. Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. Br J Surg 1991;78(3):355–60. Epub 1991/03/01. PubMed PMID: 2021856. [PubMed: 2021856]

12. Kalisvaart KJ, Vreeswijk R, de Jonghe JF, van der Ploeg T, van Gool WA, Eikelenboom P. Risk factors and prediction of postoperative delirium in elderly hip-surgery patients: implementation and validation of a medical risk factor model. J Am Geriatr Soc 2006;54(5):817–22. Epub 2006/05/16. doi: 10.1111/j.1532-5415.2006.00704.x. PubMed PMID: 16696749. [PubMed: 16696749]

13. Aldemir M, Ozen S, Kara IH, Sir A, Bac B. Predisposing factors for delirium in the surgical intensive care unit. Crit Care 2001;5(5):265–70. Epub 2001/12/12. PubMed PMID: 11737901; PubMed Central PMCID: PMCPMC83853. [PubMed: 11737901]

14. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010;48(6 Suppl):S106–13. Epub 2010/05/18. doi: 10.1097/MLR.0b013e3181de9e17. PubMed PMID: 20473190. [PubMed: 20473190]

15. van den Boogaard M, Pickkers P, Slooter AJ, Kuiper MA, Spronk PE, van der Voort PH, et al. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: observational multicentre study. Bmj 2012;344:e420 Epub 2012/02/11. doi: 10.1136/bmj.e420. PubMed PMID: 22323509; PubMed Central PMCID: PMCPMC3276486. [PubMed: 22323509]

16. Gavinski K, Carnahan R, Weckmann M. Validation of the delirium observation screening scale in a hospitalized older population. J Hosp Med 2016;11(7):494–7. Epub 2016/03/13. doi: 10.1002/jhm.2580. PubMed PMID: 26970312; PubMed Central PMCID: PMCPMC4931982. [PubMed: 26970312]

17. Faradmal J, Soltanian AR, Roshanaei G, Khodabakhshi R, Kasaeian A. Comparison of the performance of log-logistic regression and artificial neural networks for predicting breast cancer relapse. Asian Pac J Cancer Prev 2014;15(14):5883–8. Epub 2014/08/02. PubMed PMID: 25081718. [PubMed: 25081718]

18. Huang HH, Xu T, Yang J. Comparing logistic regression, support vector machines, and permanental classification methods in predicting hypertension. BMC Proc 2014;8(Suppl 1):S96 Epub 2014/12/19. doi: 10.1186/1753-6561-8-s1-s96. PubMed PMID: 25519351; PubMed Central PMCID: PMCPMC4143639. [PubMed: 25519351]

19. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications 2008;34(1):366–74. doi: 10.1016/j.eswa.2006.09.004.

2015;50(11):1641–56. Epub 2015/07/15. doi: 10.1007/s00127-015-1092-4. PubMed PMID: 26164028. [PubMed: 26164028]

37. Huang R, Moudon AV, Cook AJ, Drewnowski A. The spatial clustering of obesity: does the built environment matter? J Hum Nutr Diet 2015;28(6):604–12. Epub 2014/10/04. doi: 10.1111/jhn. 12279. PubMed PMID: 25280252. [PubMed: 25280252]

38. Shariff-Marco S, Yang J, John EM, Kurian AW, Cheng I, Leung R, et al. Intersection of Race/ Ethnicity and Socioeconomic Status in Mortality After Breast Cancer. J Community Health 2015;40(6):1287–99. Epub 2015/06/15. doi: 10.1007/s10900-015-0052-y. PubMed PMID: 26072260; PubMed Central PMCID: PMCPMC4628564. [PubMed: 26072260]

39. Radford PD, Derbyshire LF, Shalhoub J, Fitzgerald JE. Publication of surgeon specific outcome data: a review of implementation, controversies and the potential impact on surgical training. Int J Surg 2015;13:211–6. Epub 2014/12/17. doi: 10.1016/j.ijsu.2014.11.049. PubMed PMID: 25498494. [PubMed: 25498494]

40. Glance LG, Kellermann AL, Hannan EL, Fleisher LA, Eaton MP, Dutton RP, et al. The impact of anesthesiologists on coronary artery bypass graft surgery outcomes. Anesth Analg 2015;120(3): 526–33. Epub 2015/02/20. doi: 10.1213/ane.0000000000000522. PubMed PMID: 25695571. [PubMed: 25695571]
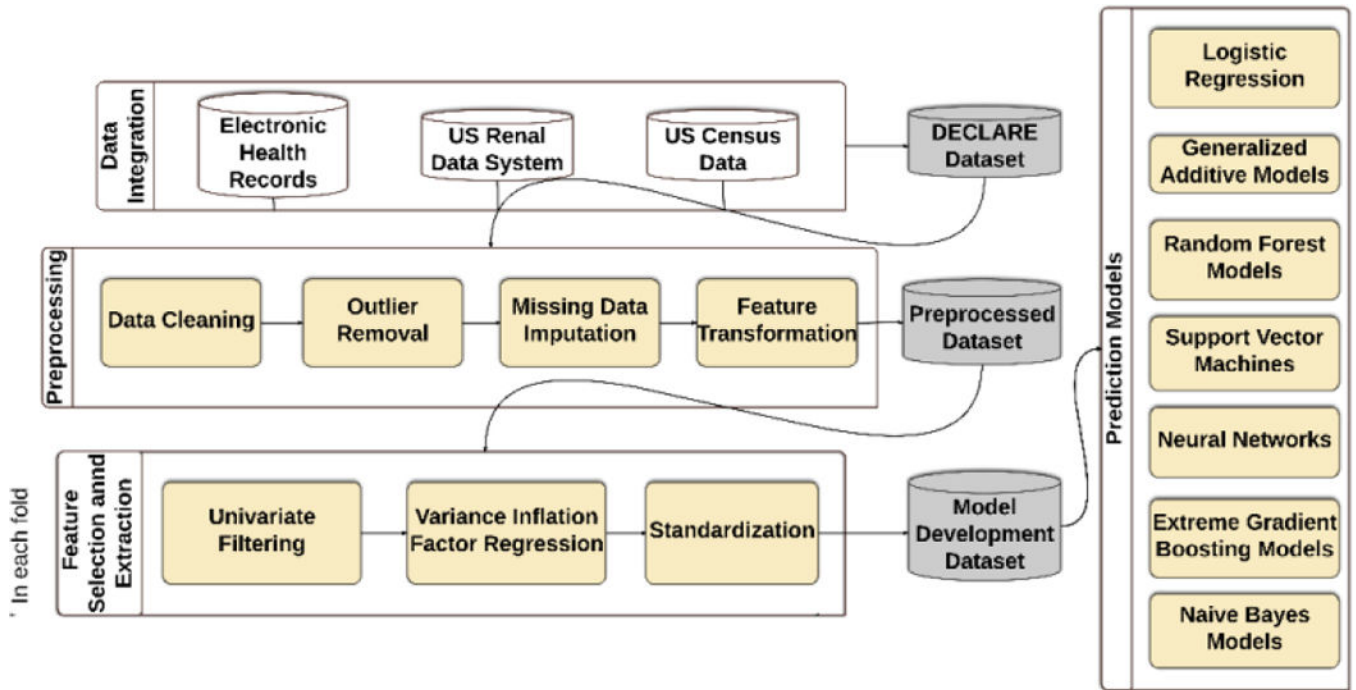
Figure 1.
The analysis workflow for prediction of delirium from EHR data.

Figure 2.
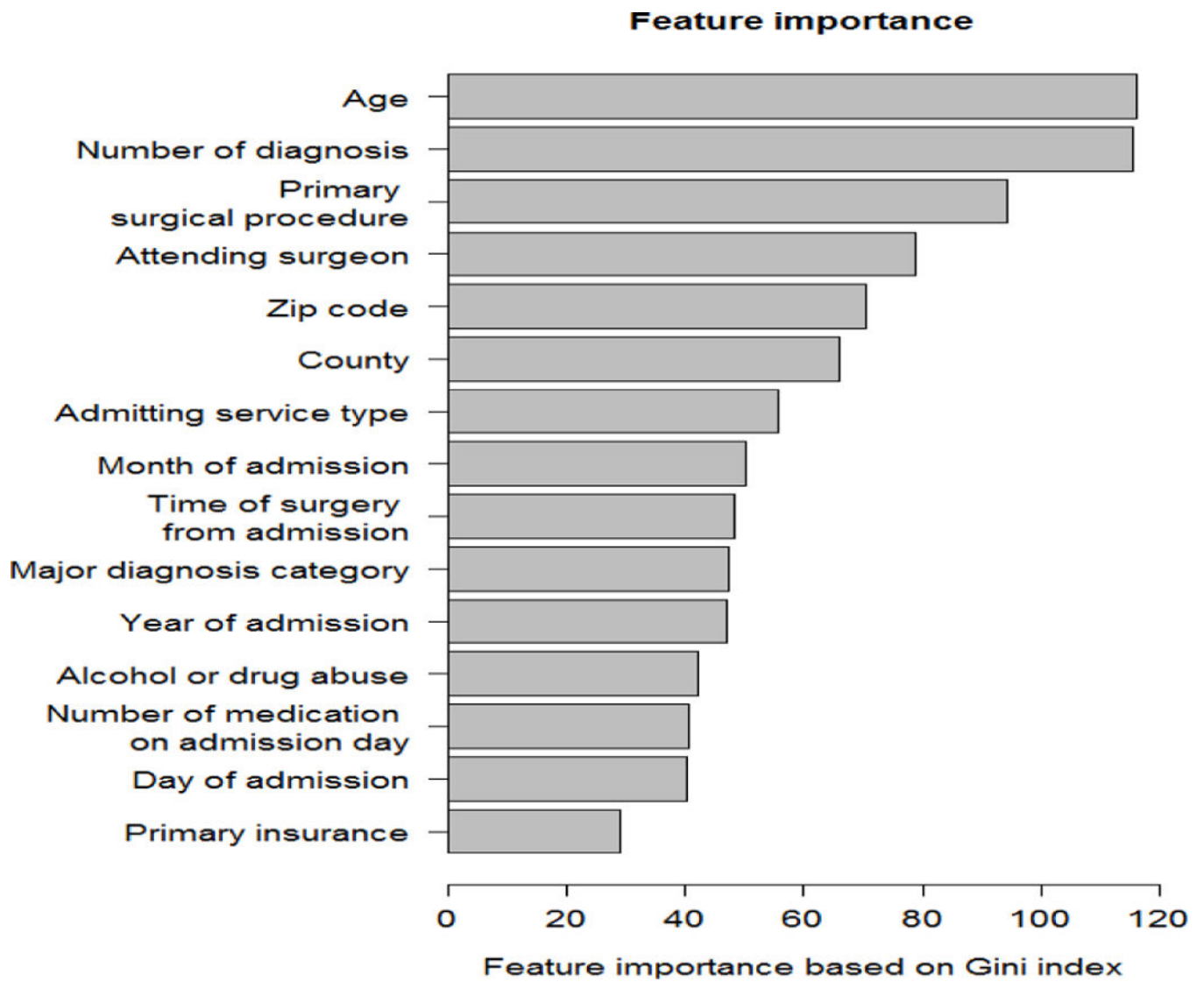Importance of top 15 features based on mean decrease in Gini impurity index for the random forests model.

Table I.

Summary of population demographics

Variables	N=51,457
Age, median (25%–75%)	56 (43, 68)
Female gender, n (%)	26337 (49)
Race, n (%)	
- White	41142 (80)
- Hispanic	6584 (13)
- African American	1600 (3)
- Other	1088 (2)

Table II.

Distribution of delirium cases among various surgery types.

Surgery type (n)	Number of subjects with delirium (%)
All (51457)	1608 (3.1)
Cardiothoracic (6890)	235 (3.4)
Non-cardiac general (20051)	313 (1.6)
Neurologic (8422)	296 (3.5)
Specialty (14740)	452 (3.1)
Other (1354) [a]	312 (23.0)

[a.] Other surgeries include ear-nose-throat, ophthalmology, and plastic surgeries.

Table III.

Prediction performance of the models, a) data, b) with downsampling, c) with SMOTE. LR: logistic regression, GAM: generalized additive model, RF: random forests, SVM: support vector machine, NN: neural network, XGB: extreme gradient boosting, NB: Naïve Bayes. AUC: Area under the ROC Curve. CI: Confidence interval.

a)

Model	Accuracy (CI 0.05, 0.95)	AUC (CI 0.05, 0.95)	Sensitivity (CI 0.05, 0.95)	Specificity (CI 0.05, 0.95)
LR	0.77 (0.68,0.83)	0.86 (0.85,0.87)	0.78 (0.70,0.88)	0.77 (0.67,0.83)
GAM	**0.81 (0.76,0.83)**	**0.86 (0.84,0.88)**	**0.75 (0.69,0.83)**	**0.81 (0.76,0.83)**
RF	**0.80 (0.75,0.87)**	**0.85 (0.83,0.86)**	**0.81 (0.75,0.88)**	**0.73 (0.58,0.80)**
SVM	0.84 (0.71,0.89)	0.71 (0.62,0.75)	0.48 (0.31,0.62)	0.85 (0.72,0.91)
NN	0.57 (0.48,0.71)	0.85 (0.83,0.87)	0.56 (0.47,0.70)	0.90 (0.81,0.95)
XGB	0.73 (0.49,0.77)	0.85 (0.84,0.86)	0.73 (0.47,0.77)	0.82 (0.76,0.96)
NB	0.75 (0.65,0.77)	0.79 (0.77,0.82)	0.71 (0.66,0.81)	0.75 (0.64,0.77)

b)

Model	Accuracy (CI 0.05, 0.95)	AUC (CI 0.05, 0.95)	Sensitivity (CI 0.05, 0.95)	Specificity (CI 0.05, 0.95)
LR	**0.81 (0.72,0.83)**	**0.86 (0.84,0.87)**	**0.74 (0.67,0.85)**	**0.82 (0.72,0.84)**
GAM	0.74 (0.72,0.75)	0.86 (0.84,0.87)	0.82 (0.79,0.86)	0.73 (0.72,0.75)
RF	**0.78 (0.76,0.80)**	**0.86 (0.84,0.87)**	**0.78 (0.75,0.80)**	**0.76 (0.70,0.81)**
SVM	0.73 (0.69,0.77)	0.85 (0.83,0.87)	0.82 (0.76,0.89)	0.71 (0.69,0.77)
NN	0.86 (0.03,0.92)	0.85 (0.83,0.87)	0.87 (0,0.94)	0.60 (0.39,1)
XGB	0.64 (0.60,0.75)	0.85 (0.83,0.86)	0.63 (0.59,0.75)	0.88 (0.79,0.92)
NB	0.68 (0.66,0.71)	0.79 (0.77,0.81)	0.79 (0.74,0.84)	0.68 (0.65,0.71)

c)

Model	Accuracy (CI 0.05, 0.95)	AUC (CI 0.05, 0.95)	Sensitivity (CI 0.05, 0.95)	Specificity (CI 0.05, 0.95)
LR	0.75 (0.73, 0.81)	0.85 (0.83, 0.86)	0.80 (0.70, 0.85)	0.75 (0.72, 0.81)
GAM	0.69 (0.62, 0.74)	0.85 (0.84, 0.87)	0.86 (0.81, 0.91)	0.68 (0.61, 0.74)
RF	**0.81 (0.68, 0.89)**	**0.85 (0.84, 0.87)**	**0.81 (0.67, 0.90)**	**0.73 (0.56, 0.85)**
SVM	0.70 (0.65,0.74)	0.84 (0.82,0.87)	0.84 (0.77,0.90)	0.69 (0.66,0.75)
NN	0.46 (0.03, 0.76)	0.84 (0.75, 0.86)	0.97 (0.77, 1)	0.44 (0, 0.76)
XGB	0.60 (0.10, 0.71)	0.84 (0.81, 0.85)	0.59 (0.07, 0.71)	0.90 (0.79, 1)

| NB | 0.79 (0.78, 0.90) | 0.80 (0.77, 0.81) | 0.75 (0.68, 0.79) | 0.71 (0.70, 0.76) |