pHealth 2021
B. Blobel and M. Giacomini (Eds.)

doi:10.3233/SHTI210578

Predicting the Aortic Aneurysm Postoperative Risks Based on Russian Integrated Data

Iuliia LENIVTCEVA[a,1], Sofia GRECHISHCHEVA[a], Georgy KOPANITSA[a] , Dmitry PANFILOV[b] and Boris KOZLOV[b]
[a] *ITMO University, 49 Kronverskiy prospect, 197101, Saint Petersburg, Russia*
[b] *Cardiology Research Institute, Tomsk National Research Medical Center of the Russian Academy of Science, Tomsk, Russia*

Abstract. This article describes the results of feature extraction from unstructured medical records and prediction of postoperative complications for patients with thoracic aortic aneurysm operations using machine learning algorithms. The datasets from two different medical centers were integrated. Seventy-two features were extracted from Russian unstructured medical records. We formulated 8 target features: Mortality, Temporary neurological deficit (TND), Permanent neurological deficit (PND), Prolonged (> 7 days) lung ventilation (LV), Renal replacement therapy (RRT), Bleeding, Myocardial infarction (MI), Multiple organ failure (MOF). XGBoost showed the best performance for most target variables (F-measure 0.74-0.95) which is comparable to recent results in cardiovascular postoperative risks prediction.

Keywords. Postoperative risks, aortic aneurysm, integrated data, predictive modeling, feature extraction, machine learning

1. Introduction

Thoracic aortic aneurysm (TAA) is a dilatation of the aorta to more than 150% of normal diameter [1] in ascending, descending aorta, or aortic arch. TAA has a 1-year mortality rate up to 75% [2]. The causes of death include not only aortic rupture, but also such complications as myocardial infarction, renal insufficiency, bleeding, stroke, etc. [3]. These risks are often compounded by several cardiovascular comorbidities which complicates the decision making. The prediction of complications is one of the ways to reduce patient's risks. Machine-learning (ML) offers an approach for risks prediction to address patient's state [4]. It uses routine clinical data to create risks prediction models. The overview of current cardiovascular postoperative models for risks prediction is shown in Table 1. About 80% essential medical data are stored in free-text medical records [5] which limits the number of data available and complicates the prediction models development.

[1] Corresponding Author, Iuliia Lenivtceva, ITMO University, 49 Kronverkskiy prospect, 197101 Saint Petersburg, Russian Federation; E-mail: lenivezzki@gmail.com.

The aim of this work is to develop a model for postoperative risks prediction for patients with TAA based on data from two Russian medical institutions concerning both structured data and free-text medical records.

Table 1. Recent algorithms, estimations, and data

Study	Algorithm	AUC-ROC	Data	Target
Lee, 2018 [6]	XGBoost	0.78	Open heart and TAA surgery	Acute kidney injury
Zhong, 2021 [7]	XGBoost	0.93	Coronary artery bypass surgery, aortic valve replacement and other heart surgeries	30-day mortality, septic shock, liver dysfunction, and thrombocytopenia
Allyn, 2017 [8]	Model ensemble	0.78	Elective heart surgery	Postoperative mortality
Fernandes, 2021 [9]	XGBoost	0.88	Intraoperative open heart surgery data	Postoperative mortality
Coulson, 2020 [10]	Logistic regression	0.78–0.85	Open heart surgery	Acute kidney injury

2. Method

2.1. Data and Features

The predictive model was developed based on two datasets. The first dataset includes structured data from Tomsk National Medical Research Center of the Russian Academy of Sciences and the second dataset includes unstructured medical records from Almazov National Medical Research Center (St. Petersburg, Russia). Tomsk dataset contains 97 structured records for 97 patients with data on aortic operations. Almazov National Medical Research Center dataset contains 56 929 text documents (2008 – 2019) for 343 TAA operations and 319 patients. We formulated 8 target features: Mortality, Temporary neurological deficit (TND), Permanent neurological deficit (PND), Prolonged (> 7 days) lung ventilation (LV), Renal replacement therapy (RRT), Bleeding, Myocardial infarction (MI), Multiple organ failure (MOF). In total, 63 input features were formulated for risks prediction. We organized these features in groups: anthropometric information (6 features), concomitant diseases (8 features), laboratory tests (5 features), coronary angiogram (4 features), echocardiography (8 features), computed tomography (14 features), intraoperative information (15 features), combined surgeries (3 features).

2.2. Feature Extraction

We extracted input and target features from Almazov National Medical Research Center text records; the Tomsk National Medical Research Center of the Russian Academy of Sciences data were already structured. All the data were anonymized by the source medical institutions. Textual data preprocessing included several steps: data cleaning, lemmatization, stop-words and rare words removal, sentences segmentation, POS-tagging, negation detection and removal, tokenization, and vectorization (TF-IDF). To realize these steps, we used the following Python packages: pymorphy2 (to work with Russian language), NLTK, spaCy. Data filtering was organized both by keywords search (for each feature) and by applying shallow algorithms: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and k-nearest neighbors (k-NN). Time frames were considered for data filtering. For instance, preoperative features were extracted from the documents before the operation. The features are extracted using the list of patterns and rules. Feature extraction accuracy was evaluated on 200 manually processed textual records. After the feature extraction step, two datasets were integrated based on feature names.

2.3. Predictive Model

Firstly, data were prepared for modelling: normalized in an interval [0,1], processed strong-correlated features (Pearson correlation), removed features with more than 60% gaps, otherwise missing values were imputed using k-NN method from sklearn package. Secondly, four algorithms were used for feature selection: univariate feature selection based on chi-squares, recursive feature elimination, decision tree ensemble and Lasso regression. Each algorithm selects 10 features and direct them to majority voting. From 8 to 11 features are selected for each target. Eight models were built to predict eight targets and a set of selected features was created for each target. Some target features can also be used as input features for other targets. We tested three algorithms for modelling: 1) LR; 2) RF; 3) XGBoost. SMOTE was used for integrated dataset balancing. The results are estimated by AUC-ROC, F-measure, and Accuracy scores, using 20-fold cross-validation. We also compared the performance of the developed models before (97 operations from Tomsk dataset) and after (440 operations, integrated dataset) extending structured dataset with extracted features.

3. Results

3.1. Data Description

Table 2 shows the percentage of missing values in extracted data.

Table 2. The percentage of missing values (only features that have missing values)

Feature	Missing, %	Feature	Missing, %
Circulatory arrest time	69.6	Cardioplegic arrest time	51.6
Cardiopulmonary bypass time	44.0	Postoperative creatinine	17.9
Aortic arch diameter	12.3	Sinuses of Valsalva diameter	12.3
Surgery duration	10.9	Blood loss	6.2
Height	4.7	Body mass index	4.7
Body surface area	4.7	Ascending aorta diameter	4.1
Left ventricle ejection fraction	3.8	Postoperative hematocrit	3.5
Weight	3.2	Age	2.9

3.2. Predictive Modelling Results

XGBoost strategy in combination with SMOTE yields the best results for most targets. Table 3 represents the best results for each target.

Table 3. The results for predictive modelling

Target	Strategy	Accuracy	AUC-ROC	F-measure
Mortality	XGBoost + SMOTE	0.915	0.928	0.872
TND	XGBoost + SMOTE	0.799	0.846	0.744
PND	XGBoost + SMOTE	0.850	0.932	0.845
Prolonged LV	XGBoost + SMOTE	0.927	0.988	0.948
RRT	XGBoost + SMOTE	0.975	0.986	0.950
Bleeding	RF + SMOTE	0.925	0.987	0.933
MI	RF + SMOTE	0.957	0.986	0.953
MOF	XGBoost + SMOTE	0.903	0.941	0.885

The comparison results before and after structuring are represented in Table 4.

Table 4. Comparing the models' performance before and after structuring textual data

Target	AUC-ROC (before)	F-measure (before)	AUC-ROC (after)	F-measure (after)
Mortality	0.845	**0.928**	0.852	**0.872**
TND	0.828	**0.846**	**0.839**	0.744
PND	0.929	**0.932**	**0.931**	0.845
Prolonged LV	0.911	**0.988**	0.909	**0.948**
RRT	0.771	**0.986**	0.784	**0.950**
Bleeding	0.893	**0.987**	0.889	**0.933**
MI	0.833	**0.987**	0.821	**0.953**
MOF	0.835	**0.941**	0.816	**0.885**

Table 5 shows top-5 most important input features for target prediction.

Table 5. Most important features for targets

Target	Top-5 important features	Target	Top-5 important features
Mortality	1) MOF 2) RRT 3) Fresh frozen plasma, units 4) Deep hypothermia 5) Age	TND	1) Entry site of aortic dissection at the sinotubular junction 2) Descending aortic dissection 3) Left internal carotid artery stenosis (50-75%) 4) Resternotomy for bleeding 5) Ascending aortic dissection
PND	1) Diameter of sinus of Valsalva 2) Entry site of aortic dissection at sinotubular junction 3) Diameter of aortic arch 4) Prolonged LV 5) Right coronary artery stenosis	Prolonged LV	1) Red blood cells, units 2) Fresh frozen plasma, units 3) Bleeding 4) Left internal carotid artery stenosis (<50%) 5) RRT
RRT	1) MOF 2) Postoperative creatinine 3) Extension of aortic dissection down to iliac and/or femoral arteries 4) Fresh frozen plasma, units 5) Prolonged LV	Bleeding	1) Previous MI 2) Aortic valve replacement 3) Height 4) Fresh frozen plasma, units 5) Retrograde dissection
MI	1) Left coronary artery stenosis (>75%) 2) Red blood cells, units 3) Drainage blood loss 4) Dissection of abdominal aorta 5) Left coronary artery stenosis (<50%)	MOF	1) RRT 2) Right coronary artery stenosis 3) Fresh frozen plasma, units 4) Red blood cells, units 5)Previous cerebrovascular accident

4. Discussion

This work is dedicated to the development of the predictive model based on the integrated medical data. For this purpose, we used two datasets from real medical institutions which contain heterogeneous data for patients with TAA operations. For integration purposes the textual data were processed to extract essential features. The extracted features were validated based on the accuracy score on the test sample. We dropped 6 features due to the differences in data formats storage, diagnostic methods for different institutions and due to the missing values. For instance, circulatory arrest time is a feature that characterizes duration of circulatory arrest in minutes, however, for Almazov National Medical Research Center data, it is often possible to extract information about circulatory arrest only as a binary feature – if a procedure was done or not. As a result of the exploratory data analysis some features such as weight (correlated with two other features), circulatory arrest time, cardioplegic arrest time, and cardiopulmonary bypass time were removed due to the large number of missing values (see Table 2) as the use of imputing techniques can affect the quality of the predictive model. To develop a predictive model three machine-learning algorithms were used: 1) logistic regression; 2) XGBoost; 3) random forest. XGBoost algorithm in combination with SMOTE showed the best results for most targets (see Table 3). It also shows comparable results to other studies in predicting postoperative cardiovascular complications (Table 1). The developed predictive model has a high potential for the thoracic aortic surgery risks prediction. Although it should be noted that from the clinical point of view the impact of several parameters in the predictive model is obscure. However, number of them has logical explanation. For example, direct relation of the aortic diameter at the sinuses of Valsalva to temporal neurological deficit is unclear. To find the answer one need to solve a logical chain. Large aortic root is an indication for its replacement. Naturally, it prolongs cardiopulmonary bypass time and, in turn, increases neurological deficit risks.

Our study has some limitations. However, we integrated data from several datasets, the number of patients and operations is relatively small and needs to be extended. We also faced with the imbalance problem during the study, which is usual for medical data [11]. In such situations machine-learning algorithms tend to classify the data into predominant class. To address this problem, we used SMOTE for data balancing and F-measure as a metric which is less sensitive to data imbalance. However, the work with imbalanced medical datasets is still an issue. One more limitation relates to the data losses during the integration process. There is a need not only to compare and map the logical data structures and contents, but also diagnostic methods and treatment approaches in different institutions as it may influence the data collected and stored. However, despite all the mentioned limitations the study showed that data structuring and integration helps to extend the dataset and improve the quality of the predictive model.

5. Conclusion

In this study we developed a model for postoperative risks prediction for patients with TAA based on data from two Russian medical institutions concerning both structured data and free-text medical records. Our study showed that heterogeneous data integration improves the performance of predictive model. Future studies may address current limitations of the study such as relevant synthetic patients' generation and model validation in a medical practice.

Acknowledgments

This work is financially supported by National Center for Cognitive Research of ITMO University.

References

[1] Capuñay C. Carrascosa P. Chapter 19 - Magnetic Resonance Angiography. In: Ţintoiu IC, Ursulescu A, Elefteriades JA, Underwood MJ, Droc I (Eds.) *New Approaches to Aortic Diseases from Valve to Abdominal Bifurcation*, Academic Press, 2018, pp. 207–214.
[2] Mohr JP, Benavente O, Barnett HJM. Chapter 16 - Spinal Cord Ischemia. In: *Stroke (Fourth Edition)*, Fourth Edi., Mohr JP, Choi DW, Grotta JC, Weir B, Wolf PA (Eds.) Philadelphia: Churchill Livingstone, 2004, pp. 423–436.
[3] Orandi BJ, Dimick JB, Deeb GM, Patel HJ, Upchurch GR. A population-based analysis of endovascular versus open thoracic aortic aneurysm repair. *J. Vasc. Surg.*, vol. 49, no. 5, pp. 1112–1116, May 2009, doi: 10.1016/j.jvs.2008.12.024.
[4] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can Machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*, vol. 12, no. 4, p. e0174944, Apr. 2017, doi: 10.1371/journal.pone.0174944.
[5] Kong HJ. Managing unstructured big data in healthcare system. *Healthc. Inform. Res.*, vol. 25, no. 1, pp. 1–2, 2019, doi: https://doi.org/10.4258/hir.2019.25.1.1.
[6] Lee H, et al. Derivation and Validation of Machine Learning Approaches to Predict Acute Kidney Injury after Cardiac Surgery. *J. Clin. Med.*, vol. 7, no. 10, p. 322, Oct. 2018, doi: 10.3390/jcm7100322.
[7] Zhong Z, Yuan X, Liu S, Yang Y, Liu F. Machine learning prediction models for prognosis of critically ill patients after open-heart surgery. *Sci. Rep.*, vol. 11, 2021, doi: 10.1038/s41598-021-83020-7.
[8] Allyn J, et al. A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: A decision curve analysis. *PLoS One*, vol. 12, no. 1, pp. 1–12, 2017, doi: 10.1371/journal.pone.0169772.
[9] Fernandes MPB, Armengol de la Hoz M, Rangasamy V, Subramaniam B. Machine Learning Models with Preoperative Risk Factors and Intraoperative Hypotension Parameters Predict Mortality After Cardiac Surgery. *J. Cardiothorac. Vasc. Anesth.*, vol. 35, no. 3, pp. 857–865, 2021, doi: 10.1053/j.jvca.2020.07.029.
[10] Coulson T, et al., Predicting Acute Kidney Injury After Cardiac Surgery Using a Simpler Model. *J. Cardiothorac. Vasc. Anesth.*, vol. 35, no. 3, pp. 866–873, 2021, doi: 10.1053/j.jvca.2020.06.072.
[11] Zhong Z, Yuan X, Liu S, Yang Y, Liu F. Machine learning prediction models for prognosis of critically ill patients after open-heart surgery. *Sci. Reports 2021 111*, vol. 11, no. 1, pp. 1–10, Feb. 2021, doi: 10.1038/s41598-021-83020-7.