Development of a VTE prediction model based on automatically selected features in glioma patients

Leontev Sergei[1][0009-0000-3967-6604], Simakova Maria[2][0000-0001-9478-1941], Lukinov Vitaly[3][0000-0002-3411-508X], Pishchulov Konstantin[2][0000-0003-3575-3945], Derevitskii Ilia[1][0000-0002-8624-5046], Abramyan Levon[2][0000-0002-6830-5172], Vatian Alexandra[1][0000-0002-5483-716X]

[1] ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia, international@itmo.ru

[2] Personalized Medicine Centre, Almazov National Medical Research Centre, Saint Petersburg, Russian Federation

[3] Institute of Computational Mathematics and Mathematical Geophysics SB RAS

Abstract. Venous thromboembolism (VTE) poses a significant risk to patients undergoing cancer treatment, particularly in the context of advanced and metastatic disease. In the realm of neuro-oncology, the incidence of VTE varies depending on tumor location and stage, with certain primary and secondary brain tumors exhibiting a higher propensity for thrombotic events. In this study, we employ advanced machine learning techniques, specifically XGBoost, to develop identifying models for predictors searching associated with VTE risk in patients with gliomas. By comparing the diagnosis testing accuracy of our XGBoost models with traditional logistic regression approaches, we aim to enhance our understanding of VTE prediction in this population. Our findings contribute to the growing body of literature on thrombosis risk assessment in cancer patients and may inform the development of personalized prevention and treatment strategies to mitigate the burden of VTE in individuals with gliomas at the hospital term.

Keywords: VTE, Predictive Modeling, Machine Learning, Clinical diagnosis.

1 Introduction

Venous thromboembolism (VTE) is a condition that encompasses superficial vein thrombosis, deep vein thrombosis (DVT), venous gangrene, and pulmonary embolism (PE). In the field of cardio-oncology, cancer-associated thrombosis is a significant concern and is strongly associated with increased early all-cause mortality during cancer chemotherapy and surgery [1, 2, 3].

The construction of VTE prediction models remains an urgent issue to this day [4, 5, 6, 7, 8].

It has been observed that certain cancer sites such as the pancreas, kidneys, ovaries, lungs, gastrointestinal tract, and brain tumors have a higher tendency to cause blood clots. These tumors are categorized as either primary or secondary tumors that are linked with metastasis. The two most prevalent primary brain tumors are meningioma and glial tumors, which account for 35,6% and 35,5% of cases, respectively [9].

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_34

The purpose of this study was to compare XGBoost and logistic regression methods as tools for creating a risk stratification model for venous thromboembolic events in patients with gliomas.

2 Materials and methods

A study was conducted at the Almazov National Medical Research Center from January 2021 to May 2023, which enrolled 286 consecutive patients with histologically verified glioma who underwent surgery. The group consisted of 133 (51,2%) men and 132 (49,8%) women, with an average age of 54 [41; 63] years. The diagnosis of pulmonary embolism and deep vein thrombosis was made in accordance with current clinical recommendations [10].

The study determined the frequency of binary variables indicating the occurrence of VTE, clinical manifestations of neoplasms, and concomitant cardiovascular pathology. 95% confidence intervals (95% CI) were estimated using the Wilson formula and compared by the Fisher’s exact test.

The models were compared by the areas under the ROC curves (AUC) using the DeLong test.

Prognostic characteristics related to VTE development were also compared. Sensitivity and specificity were assessed using McNemar's test, while positive and negative predictive values (PPV and NPV) were compared using a weighted generalized test (WSG test).

To address the issue of multiple comparisons, p-values were adjusted using the Benjamini-Hochberg method. Statistical hypotheses were tested at a significance level of p = 0,05. Differences were considered statistically significant if p < 0,05.

3 Model Building

3.1 Metrics

The parameters chosen for maximizing in model construction are specificity and precision (sensitivity). This means that models with higher sum of specificity and precision will be considered of higher quality.

3.2 Feature selection for the final model

XGBoost models were chosen both as the final model and the feature selection model. These models were partitioned into test and training sets based on the rules described in previous chapters.

Subsequently, the remaining parameters underwent sequential inclusion into the model. Those parameters which led to the greatest increase in specificity were chosen. The process involved starting with one parameter and then adding another, continuing until sets of parameters ranging from 5 to 10 pieces were studied. The final model exhibited the most favorable results with 7 parameters.

The selected parameters were subsequently automatically transferred to the main model for its training.

3.3 Model building

Next, the final model was built, which dynamically receives as input the parameters obtained in previous stage. The model itself is also an XGBoost, but with parameters different from those that were used when selecting parameters.

3.4 Results

The parameter selection stage for the transmitted data resulted in the following 7 parameters: ['D-dimer', 'BMI', 'bed rest (more than 3 days), prolonged lying position', 'PulmonaryDis', 'varicose veins', 'Hypertension', 'Dyslipidemia'].

The model was able to achieve the following indicators according to the main metrics (Table 1).

Table 1. Model result metrics.

Metric	Values [95% CI]
Specificity	93% [87%; 99%]
Precision	77% [58%; 94%]
Accuracy	84% [76%; 90%]
Recall	61% [42%; 77%]
F1-score	68% [51%; 81%]

Diagnostic testing accuracy table is shown in Fig. 1.

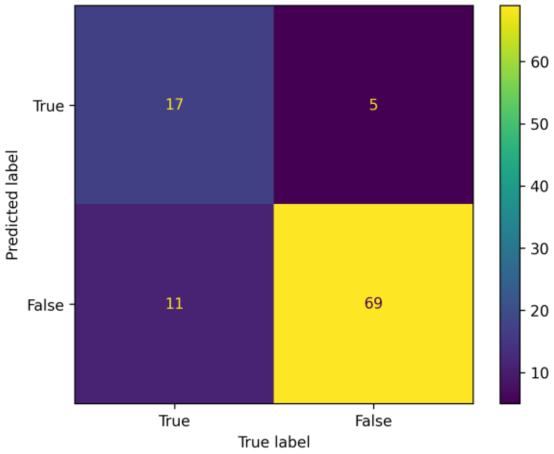

Fig. 1. Diagnostic testing accuracy table.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_34

The area under ROC curve is 0,77 (Fig. 2).

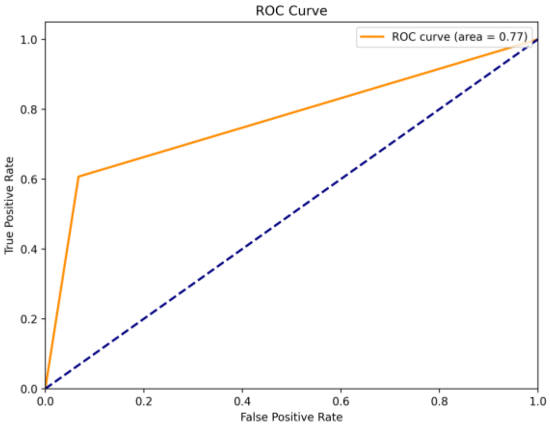

Fig. 2. ROC curve of XGBoost model.

3.5 Comparison with logistic regression

Significant predictors of VTE were identified through the construction of single-factor logistic regression models. Independent predictors of VTE development were identified through the construction of a multi-factor logistic regression model. The data from both single-factor and multi-factor regression analyses are presented in Table 2.

Table 2. The data from the multi-factor and single-factor regression analyses.

Covariates	Single-factor models p	Multi-factor models p
bed rest (more than 3 days), prolonged lying position	<0,001*	<0,001*
D-dimer	<0,001*	0,006*
PLT	0,006*	0,099
Age at the time of inclusion	0,010*	0,067
Radiation therapy	0,044*	0,047*

The summary characteristic of the model based on ROC analysis data is presented in the Table 3 and on Fig. 3.

Table 3. The summary characteristic of the logistic regression.

Parameter	Values [95% CI]
Specificity	78,6% [49,2%; 95,3%]
Sensitivity	93,5% [88,4%; 96,8%]

Positive Predictive Value	52,4% [29,8%; 74,3%]
Negative Predictive Value	98% [94,2%; 99,6%]
Positive Likelihood Ratio	12,1 [6,3; 23,4]
Negative Likelihood Ratio	0,2 [0,1; 0,6]

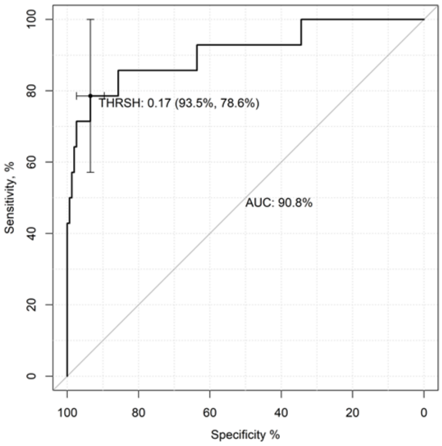

Fig. 3. Roc curve of multifactor logistic regression model.

The next step involved validating the model on a prospective sample of 100 patients with CNS gliomas who underwent treatment at the V.A. Almazov National Medical Research Center of the Ministry of Health of Russia during the period from 2022 to 2023. The validation data is presented in Table 4.

Table 4. The internal validation data of the model on the prospective sample.

Parameter	Values [95% CI]
Specificity	95% [87%; 99%]
Sensitivity	47% [23%; 72%]
Positive Predictive Value	73% [39%; 94%]
Negative Predictive Value	85% [77%; 94%]
Positive Likelihood Ratio	10,2 [3,03; 34,35]
Negative Likelihood Ratio	0,56 [0,35; 0,87]

4 Model comparison

A comparison of the models in the prospective sample is presented in tables 5-8 and Fig. 4. The difference in the total number of patients is due to incomplete data.

Table 5. Diagnostic testing accuracy table of XGBoost model for all data

	Outcome +	Outcome -	Total
Test +	11	5	16
Test -	7	69	76
Total	18	74	92

Table 6. Diagnostic Testing Accuracy Table of XGBoost model for adjusted with multi-factor logistic regression data

	Outcome +	Outcome -	Total
Test +	10	5	15
Test -	7	60	67
Total	17	65	82

Table 7. Diagnostic Testing Accuracy Table of multifactor logistic regression model

	Outcome +	Outcome -	Total
Test +	8	3	11
Test -	9	65	71
Total	17	65	82

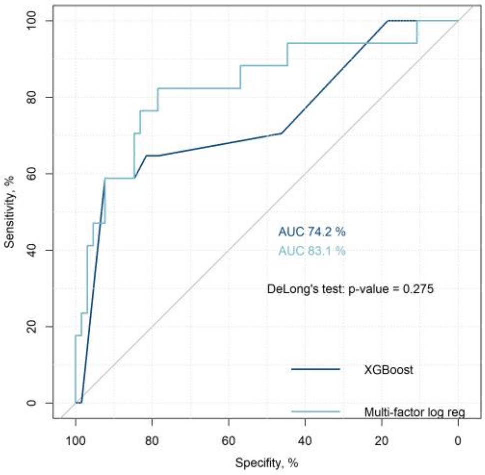

Fig. 4. Comparing ROC curves of models on prospective data.

Table 8. Comparing diagnostic models accuracy

Name	1. XGBoost model on primary prospective data	2. XGBoost model on adjusted data	3. Multi-factor logistic regression on adjusted data	Compare 1-3 2-3
	Value [95% CI]	Value [95% CI]	Value [95% CI]	P (the same)
Apparent prevalence	0,17[0,10; 0,27]	0,18[0,11; 0,28]	0,13[0,07; 0,23]	0,423
True prevalence	0,20[0,12; 0,29]	0,21[0,13; 0,31]	0,21[0,13; 0,31]	-
Sensitivity	0,61[0,36; 0,83]	0,59[0,33; 0,82]	0,47[0,23; 0,72]	0,414
Specificity	0,93[0,85; 0,98]	0,92[0,83; 0,97]	0,95[0,87; 0,99]	0,480
Positive predictive value	0,69[0,41; 0,89]	0,67[0,38; 0,88]	0,73[0,39; 0,94]	0,711
Negative predictive value	0,91[0,82; 0,96]	-	0,87[0,77; 0,94]	0,483

No statistically significant difference in the characteristics of the models was found.

5 Conclusions

In this study, a XGBoost model was constructed to predict the development of venous thromboembolism (VTE) in glioma patients. This model was based on the analysis of automatically selected parameters. An XGBoost algorithm was employed to build the model, optimized based on the obtained parameters.

The XGBoost model demonstrated good performance according to key metrics including specificity, precision, recall, and F1-score. Error analysis and ROC curve were utilized to assess the model's quality, and shap values were generated to illustrate parameter importance.

Additionally, the model was compared with logistic regression model, revealing significant predictors of VTE development. Validation of the model on an independent sample of patients confirmed its ability to generalize to external data.

Acknowledgements

This work was supported by Russian Science Foundation, Grant № 23-11-00346.

Referencses

1. Nicholson, M., Chan, N., Bhagirath, V., Ginsberg, J.: Prevention of venous thromboembolism in 2020 and beyond. Journal of Clinical Medicine 9, 1-27 (2020). doi:10.3390/jcm9082467
2. Kearon, C., Akl, E.A., Ornelas, J., Blaivas, A., Jimenez, D., Bounameaux, H., Huisman, M., King, C.S., Morris, T.A., Sood, N., et al.: Antithrombotic therapy for VTE disease: CHEST guideline and expert panel report. Chest 149 (2), 315–352 (2016). doi:10.1016/j.chest.2015.11.026
3. Connors, J.M., Levy, J.H.: COVID-19 and its implications for thrombosis and anticoagulation. Blood 135, 2033-2040 (2020). doi:10.1182/BLOOD.2020006000
4. Xu, Q., Lei, H., Li, X., Li, F., Shi, H., Wang, G., Sun, A., Wang, Y., Peng, B.: Machine learning predicts cancer-associated venous thromboembolism using clinically available variables in gastric cancer patients. Heliyon 9 (1), (2023). doi:10.1016/j.heliyon.2022.e12681
5. He, L., Luo, L., Hou, X., Liao, D., Liu, R., Ouyang, C., Wang, G.: Predicting venous thromboembolism in hospitalized trauma patients: a combination of the Caprini score and data-driven machine learning model. BMC Emergency Medicine 21 (1), (2021). doi:10.1186/s12873-021-00447-x
6. Lin, C.C., Chen, C.C., Li, C.I., Liu, C.S., Lin, W.Y., Lin, C.H., Yang, S.Y., Li, T.C.: Derivation and validation of a clinical prediction model for risks of venous thromboembolism in diabetic and general populations. Medicine (United States) 100 (39), E27367 (2021). doi:10.1097/MD.0000000000027367
7. Gerotziafas, G. T., Papageorgiou, L., Salta, S., Nikolopoulou, K., Elalamy, I.: Updated clinical models for VTE prediction in hospitalized medical patients. Thrombosis Research 164, S62–S69 (2018). doi:10.1016/j.thromres.2018.02.004
8. Beal, E.W., Tumin, D., Chakedis, J., Porter, E., Moris, D., Zhang, X. feng, Abdel-Misih, S., Dillhoff, M., Manilchuk, A., Cloyd, J., et al.: Identification of patients at high risk for post-discharge venous thromboembolism after hepato-pancreato-biliary surgery: which patients benefit from extended thromboprophylaxis? HPB 20 (7), 621–630 (2018). doi:10.1016/j.hpb.2018.01.004
9. Lee, E. J., Chang, C. H., Wang, L. C., Hung, Y. C., & Chen, H. H.: Two primary brain tumors, meningioma and glioblastoma multiforme, in opposite hemispheres of the same patient. Journal of Clinical Neuroscience 9(5), 589-591 (2002). doi: 10.1054/jocn.2002.1086. PMID: 12383424.
10. Farge, Dominique, Corinne Frere, Jean M. Connors, Alok A. Khorana, Ajay Kakkar, Cihan Ay, Andres Muñoz, et al.: 2022 international clinical practice guidelines for the treatment and prophylaxis of venous thromboembolism in patients with cancer, including patients with COVID-19. The Lancet Oncology 23, e334–e347 (2022). doi:10.1016/S1470-2045(22)00160-7