J. Mantas et al. (Eds.)

doi:10.3233/SHTI210882

Length of Stay Prediction in Neurosurgery with Russian GPT-3 Language Model Compared to Human Expectations

Gleb DANILOV[a,1], Konstantin KOTIK[a], Elena SHEVCHENKO[a], Dmitriy USACHEV[a], Michael SHIFRIN[a], Yulia STRUNINA[a], Tatyana TSUKANOVA[a], Timur ISHANKULOV[a], Vasiliy LUKSHIN[a] and Alexander POTAPOV[a]

[a] *Laboratory of Biomedical Informatics and Artificial Intelligence, National Medical Research Center for Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation*

Abstract. Patients, relatives, doctors, and healthcare providers anticipate the evidence-based length of stay (LOS) prediction in neurosurgery. This study aimed to assess the quality of LOS prediction with the GPT3 language model upon the narrative medical records in neurosurgery comparing to doctors' and patients' expectations. We found no significant difference ($p = 0.109$) between doctors', patients', and model's predictions with neurosurgeons tending to be more accurate in prognosis. The modern neural network language models demonstrate feasibility in LOS prediction.

Keywords. Length of stay, neurosurgery, prediction, neural networks, deep learning, natural language processing

1. Introduction

Patients and relatives anticipate the evidence-based risk assessment and length of stay (LOS) prediction in high-tech surgery [1,2]. LOS prognosis can also be utilized in clinical resource management. This paper continues a series of our publications on LOS predicting in neurosurgery based on unstructured textual data [5,6]. The current study aimed to assess the quality of LOS prediction with the GPT3 language model upon the narrative medical records in neurosurgery comparing to doctors' and patients' expectations.

2. Methods

Our study consisted of two components: 1) training a neural network language model to predict LOS based on unstructured text data collected retrospectively from electronic health records (EHR); 2) comparing the model predictions with the expectations of

[1] Corresponding Author, Gleb Danilov, N.N. Burdenko Neurosurgery Center, 4th Tverskaya-Yamskaya str. 16, Moscow 125047, Russian Federation; E-mail: glebda@yandex.ru.

neurosurgeons and patients on admission (before surgery) collected in prospective research.

The narrative textual data from the EHR of N.N. Burdenko National Medical Research Center of Neurosurgery was obtained for the period from 2000 to 2017 to shape the training corpus. The texts were extracted from the following source documents created upon patient admission: initial diagnosis, complaints, life history, history of present illness, somatic status, mental status, neurological examination, comorbidity, allergies, primary nursing examination, complications. Additionally, we added the narrative operative reports (for surgery commonly performed 1-2 days after admission) to the training set of texts. The omissions in all text records were replaced with blank lines. The texts from all document sources were concatenated through the ".\n" symbol into one field for each treatment case that served as the model's input. We replaced all contiguous zero-width spaces with an empty string, line-breaking spaces with a single newline, non-breaking spaces with a single space, then stripped any leading/trailing whitespace. Texts were converted to lower case. The target variable we used for machine learning was ln(1 + LOS), where LOS was the total number of days a patient stayed at the hospital since admission.

The underlying model (Generative Pre-trained Transformer 3 trained with large (600+ GB) corpus in the Russian language – ruGPT3) and tokenizer we utilized can be found here [3]. It was chosen as the promising state-of-the-art technology in Russian language modeling. The model vocabulary size was equal to 50 257 tokens. The length of the input text was limited to 2048 tokens. This, in turn, was also the maximum context length. The dimension of a token's vector representation was 768. To obtain the final vector representation of the token, the corresponding vectors were extracted from all 12 encoder layers and concatenated into one vector. To get a vector representation of the entire sentence, the final vector representations of its tokens were averaged into one vector dimensioned 768x12, which served the input to the fully connected layer (FCL) preceded by a dropout with a probability of 0.3.

The model fine-tuning was performed in two stages: 1) the top layer (FC) was trained during two epochs, the rest of the weights were "frozen"; 2) the whole model was trained during 20 epochs. The neural network training parameters were as follows: batch_size_top = 256 (training the top layer), learning_rate_top = 1e-5, batch_size_all = 16 (training the whole model), learning_rate_all = 1e-5, loss function: L1LOSS (https://pytorch.org/docs/stable/generated/torch.nn.L1Loss.html), optimizer: Adam (https://pytorch.org/docs/master/generated/torch.optim.Adam.html). The 25% random cases from 90 685 were used to evaluate prediction quality when fine-tuning the model. The inverse of the logarithm operation was applied to the model predictions to get the prognosis in days.

In a prospective study (carried out in the first half of May 2021), patients admitted to the N.N. Burdenko Neurosurgery Center and their treating neurosurgeons were asked to predict the total length of inpatient stay. Simultaneously, we registered neural model predictions.

Mean Absolute Error (MAE) in days was primarily used to measure the quality of model prediction and the correctness of doctors' and patients' expectations in a prospective study. Additionally, we calculated median absolute error (AE) and AE interquartile range (IQR) to better characterize AE distributions from different types of responders. We tested the difference in AE of prediction with the Kruskal-Wallis test. Spearman correlation coefficient was used to estimate the correlation between patients', doctors', and model's predictions. A p-value < 0.05 was considered statistically significant.

3. Results

A total of 90 688 cases of neurosurgical treatment were identified retrospectively to provide unstructured medical texts for model training. A total of 111 patients with data collected prospectively were enrolled in the testing study. The study patients (adults with average age 52.8 ± 15.0 years; males / females = 49 / 62) underwent surgery for brain tumors (n = 78), vascular pathology (n = 12), hydrocephalus (n = 5) and other (n = 16). The length of the preoperative period in 86 (77.5%) of patients in the testing sample did not exceed 2 days (median = 1). The average LOS for the studied group was 7.5 days (median – 7 [6;9] days). The maximum reached 21 days in a complicated case. The LOS prognosis MAE and median AE with IQR obtained from doctors, patients, and the model are presented in Table 1. There was no statistically significant difference in the AE of prediction between the three types of responders (p = 0.109). However, doctors' prognoses tended to be more accurate.

Table 1. MAE – mean absolute error, AE – absolute error, IQR - interquartile range

Responders	MAE	Median AE [IQR]
Neurosurgeons	1.88	1.00 [1.00, 3.00]
Patients	2.37	2.00 [1.00, 3.00]
Prognostic model	2.37	2.00 [1.00, 3.00]

We observed a weak and statistically significant correlation between doctors' and patients' expectations (rho = 0.30, p = 0.002), patients' and model's prognosis (rho = 0.31, p = 0.001) and doctors' and model's prediction (rho = 0.29, p = 0.002). A boxplot of AE distributions for each type of responders with the estimation of statistically significant differences is shown in Figure 1.

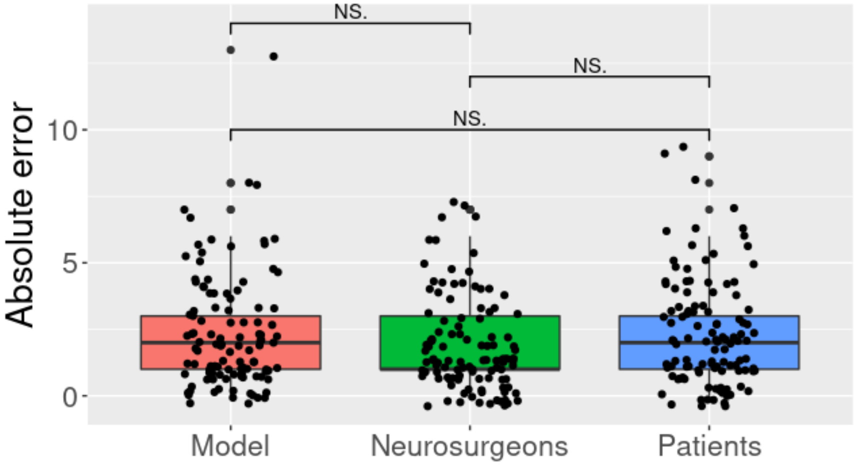

Figure 1. The distribution of prediction AE from three types of responders (the model, neurosurgeons and patients). NS. - non-significant.

4. Discussion

The length of stay is an important metric for healthcare providers and patients, and its prediction is sought to be automatized [4]. This study focused on utilizing narrative clinical text data, which seems informative and underlooked. We admit that our model had an advantage over the doctors and patients since it benefitted from the operative reports. Thus, the model, in fact, gave a prognosis a little later. However, given that surgery was performed 1-2 days after admission in most cases, the model's prognosis can still be considered timely. We also found that the neurosurgeons tend to predict LOS better than the patient or model. Nevertheless, the quality of the ruGPT3 model's prediction demonstrated in our research (even accounting for some unfairness of the comparison to humans) justifies considering such solutions as potentially applicable in automating LOS prediction. The testing sample size and its possible heterogeneity might be the limitations of the current study.

5. Conclusion

We found the medical texts sufficiently informative to solve complex tasks with modern neural network language models. The approach we justify should be considered in addition to predictive modeling based on structured feature space.

Acknowledgements

This project was supported by the Russian Foundation for Basic Research grants 18-29-01052 (information extraction) and 18-29-22085 (predictive modeling).

References

[1] Tanzer D, CRA KS, Tanzer M. Changing Patient Expectations Decreases Length of Stay in an Enhanced Recovery Program for THA. Clin. Orthop. Relat. Res. 2018;476(2):372-378. doi:10.1007/S11999.0000000000000043.

[2] Auer C, Laferton J, Shedden-Mora M, Salzmann S, Moosdorf R, Rief W. Optimizing preoperative expectations leads to a shorter length of hospital stay in CABG patients: Further results of the randomized controlled PSY-HEART trial. J. Psychosom. Res. 2017;97:82-89. doi:10.1016/J.JPSYCHORES.2017.04.008.

[3] sberbank-ai/rugpt3small_based_on_gpt2 Hugging Face, (n.d.). https://huggingface.co/sberbank-ai/rugpt3small_based_on_gpt2 (accessed August 29, 2021).

[4] Awad A, Bader-El-Den M, Mcnicholas J. Patient length of stay and mortality prediction: A survey. Heal. Serv. Manag. Res. 2017;30(2):105–120. doi:10.1177/0951484817696212.

[5] Danilov G, Kotik K, Shifrin M, Strunina U, Pronkina T, Potapov A. Prediction of Postoperative Hospital Stay with Deep Learning Based on 101 654 Operative Reports in Neurosurgery. Stud. Health Technol. Inform. 2019;258:125–129. http://www.ncbi.nlm.nih.gov/pubmed/30942728.

[6] Danilov G, Kotik K, Shifrin M, Strunina U, Pronkina T, Potapov A. Predicting Postoperative Hospital Stay in Neurosurgery with Recurrent Neural Networks Based on Operative Reports. Stud. Health Technol. Inform. 2020;270:382–386. doi:10.3233/SHTI200187.