Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200187

Predicting Postoperative Hospital Stay in Neurosurgery with Recurrent Neural Networks Based on Operative Reports

Gleb DANILOV[a,1], Konstantin KOTIK[a], Michael SHIFRIN[a], Uliya STRUNINA[a], Tatyana PRONKINA[a] and Alexander POTAPOV[a]

[a] *National Medical Research Center for Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation*

Abstract. This study aimed to predict the duration of the postoperative in-hospital period in neurosurgery based on unstructured operative reports, natural language processing, and deep learning. The recurrent neuronal network (RNN-GRU) was tuned on the word-embedded reports of primary surgical cases retrieved for the period between 2000 and 2017. A new test dataset obtained for the primary operations performed in 2018-2019 was used to evaluate model performance. The mean absolute error of prediction in the final test was 3.00 days. Our study demonstrated the usability of textual EHRs data for the prediction of postoperative period length in neurosurgery using deep learning.

Keywords. Neurosurgery, Electronic Health Records, Operative Report, Machine Learning, Deep Learning, Recurrent Neuronal Networks, Artificial Intelligence

1. Introduction

Although machine learning is gaining special attraction in neurosurgery during the last years, the text analysis and natural language processing have rarely been used [1], [2]. The operative report is a document describing the details of surgical procedures in a patient's medical records. Written narratively in the majority of clinics, it undoubtedly provides certain knowledge when appropriately explored.

In-hospital stay duration is an indirect estimate of disease severity and healthcare expenses. We hypothesized that the information stored in operative reports could be sufficiently meaningful to predict the duration of the postoperative in-hospital period. To test that hypothesis in a pilot study the predictive model (recurrent neuronal network (RNN) - gated recurrent unit (GRU)) was previously built on 75 531 and tested on 26 123 operative reports [2]. We have shown that the postoperative period could be (to a certain extent) reasonably predicted using the textual description of the surgery, natural language processing, and deep learning. However, our study was limited by the data collected in a period between 2000 and 2017 years [2]. This study aimed to update the

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

model developed previously and validate it on a new set of operative reports never seen by the model before.

2. Methods

In our previous study, the EHRs of N.N. Burdenko Neurosurgery Center was queried to select all operative reports with the corresponded length of in-hospital stay following surgery in the period between 2000 and 2017 [2]. In February 2018 the Hospital information system at N.N. Burdenko Neurosurgery Center was changed. The data on operative reports for a period between January 2018 and June 2019 were obtained from two sources: the old EHR (named "e-Med"; January - February 2018) and the new one (named "Asclepius"; February 2018 – June 2019) [3]. All the incomplete reports and cases with an inaccessible length of in-hospital stay were excluded.

In the current study, all the operative reports dated to 2000-2017 were split into primary surgical cases (the operations performed at our Center for the first time for patients) and repeated surgery, and the training samples were randomly generated from 75% of each subset. The testing samples constituted of the remaining 25% of documents in each category. The model was trained on primary and repeated surgery subsets independently, and its performance was compared in two testing samples accordingly. Finally, the model built on primary cases was fine-tested on operative reports for primary operations dated to January 2018 - June 2019. We did not consider the cases of repeated surgery for the final test due to the reasons explained in section 3.1.

2.1 Text preprocessing

All neurosurgical operative reports typed by neurosurgeons on computer keyboards were retrieved from EHRs. Corpus preparation was done using R programming environment (version 3.5.0) in RStudio IDE for MacOS (version 1.1.453). Raw textual data were preprocessed before fed into deep learning procedure as input:

- Transformed to lower case
- Tokenized with a space separator
- Proceeded by word embeddings

All the tokens generated from the original texts were mapped into numeric vectors from Euclidean space R^d when $d = 400$. These vectors were assumed to capture hidden information about a language, like word analogies or semantic relations between them. Distributed word representation was obtained from operative reports using the FastText library [4], [5]. The idea of that approach was in learning to predict a word from its context or vice versa. We considered the basic vector representations of character trigrams and expanded each word as the sum of the vectors of its trigrams. Then we built the constructions of skip-gram (a set of tokens that imply other tokens between them in an amount not exceeding the preset number) on these sums. Unlike the classic Word2vec and Glove, FastText provides the invariance of embedded tokens to word morphology since N-grams of characters are more common than whole words. As a result, the vector representations we obtained were expected to work efficiently with the rich-in-morphology Russian language.

2.2 The model architecture

The target variable was the postoperative period duration (number of days stored in EHR as integers). To process the operative reports, we RNN-GRU architecture proposed in our previous study [2], [6]. The logarithm of the hyperbolic cosine of the prediction error was used as a loss function since it is resistant to outliers and the second derivative of it exists. The adaptive gradient method RmsProp was utilized as the optimizer with the cut-off gradient value of the maximum norm equal to 1, and the cut-off gradient value outlying the range [-1,1] [4]. The activation function used was Exponential Linear Unit (ELU) [4]. The hyperparameters of the model were set by the grid search as:

- Batch size = 256 texts
- Dropout (for regularization) = 0.3

The model was trained using a Keras framework (keras.io) on the first 500 vector-represented tokens corresponding to documents. When the number of tokens was < 500, the sequence was padded with a fixed token complementing sequence to 500.

2.3 The evaluation of model performance

Mean Absolute Error (MAE) was used to measure the quality of model prediction:

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j| \quad (1)$$

where n – was the number of operative reports, y_+ – true length of the postoperative in-hospital period, $\hat{y}_+$ – the predicted duration of postoperative in-hospital stay [7].

The model was trained and evaluated within the Python (version 3.6) environment Jupyter Notebook for MacOS.

3. Results

In the period between 2000 and 2017, a total of 77 876 patients (avg. age 39 ± 20 years, males – 55%) underwent 104 506 neurosurgical operations. Initially, we obtained the complete uncorrupted operative reports for 101 664 operations and could verify the duration of postoperative period in 101 654 cases eligible for the inclusion. These texts consisted of 163 316 unique words or 36 438 unique tokens excluding stop-words and words appeared five and fewer times in all reports. The "primary surgery" subset included 75 652 reports (of which 56 739 documents were selected as a training sample). The "repeated surgery" subset included 26 002 reports (with 19 501 used for training).

3.1 Training and validation of the models

The characteristics of samples subsetted for this study and the corresponding measures of deep learning performance are presented in Table 1. The RNN-GRU was applied to "primary surgery" and "repeated surgery" datasets, resulted in MAE of 4.83 and 14.25 days respectively. The model trained on the first dataset outperformed the model previously built on the whole set of data for the period 2000-2017 (Table 1, primary

surgery vs. all reports). However, the result of prediction in the "repeated surgery" dataset appeared to be three times worse. Thus, the RNN-GRU built on the primary surgery dataset was considered as the final model.

3.2 Testing the final model

A total of 2034 operative reports for neurosurgical procedures was obtained from "e-Med" and 11 460 reports were retrieved from "Asclepius" for the period between January 2018 and June 2019. The test of a final model was performed on 11 968 complete operative reports for primary surgery with the known length of the postoperative period (a new testing dataset). The final model demonstrated MAE = 3.00 days (Table 1). The distinct result for "e-Med" data (n = 1 829, MAE = 2.85 days) appeared to be similar for that from "Asclepius" (n = 10 139, MAE = 3.03 days).

Table 1. The results of RNN (bidirectional GRU) testing on operative reports for primary and repeated surgery (n – number of operative reports, AE – absolute error).

	All reports* (2000-2017)	Primary surgery (2000-2017)	Repeated surgery (2000-2017)	Primary surgery** (2018-2019)
Sample size, n	101 654	75 652	26 002	11 968
Range of postoperative stay, days	0 - 1251	0 - 1251	0 - 746	0 - 199
Training sample size, n	75 531	56 739	19 501	11 968
MAE, days	***7.78***	***4.83***	***14.25***	***3.00***
Predictions with AE = 0 days	14.1%	20.2%	7.7%	20.1%
Predictions with AE ≤ 2 days	54.3%	64.9%	36.3%	67.5%
Predictions with AE ≤ 3 days	***63.9%***	***73.7%***	***46.8%***	***78.5%***
Median AE, days	2.19	1.52	3.86	1.51
Maximum AE, days	1070.51	731.37	730	174.52

* the results obtained from our previous study [2]
** a new testing dataset containing operative reports for primary surgery performed in Jan 2018 - June 2019

The absolute prediction error in the new testing dataset did not exceed three days in 78.5% of cases.

4. Discussion

To the best of our knowledge, this is the first study in neurosurgery aimed at prediction of the postoperative period based on unstructured medical texts. We chose the operative report for a pilot study to test the utility of textual records. A real-world model production, however, should consider the maximum structured and unstructured data available.

W. Muhlestein et al. (2017, 2018) predicted the length of in-hospital stay after brain tumor surgery using machine learning on non-textual data [8], [9]. Non-elective surgery, preoperative pneumonia, sodium abnormality, or weight loss, and non-White race were found the strong predictors of the prolonged in-hospital stay [8], [9]. In contrast, the results we obtained demonstrate the meaningfulness of narrative medical texts in these non-trivial predictions. The worse predictive capability of repeated surgery reports may indicate that the information essential for prognosis was not captured compared to primary surgery descriptions.

The lower MAE shown in the new testing dataset could be related to the shorter range of the target variable (compared to the development datasets). The high range of AE in all samples was caused by the broad initial range of postoperative periods due to outliers. The outliers in the training samples appeared to be extremely rare and better explained by occasional factors not related to the initial surgery directly, such as severe complications and repeated interventions. The EHR software change was not supposed to affect the results since the operative reports remained in textual format invariant to the system's interface.

The RNN-GRU model was efficient to predict the postoperative period with an acceptable error for the pilot study. However, the rationale for its efficiency should be unraveled in more details, e.g., applying the other machine learning methods, combining it with structured data analysis and considering neurosurgeon's expectation of a postoperative period length. Nevertheless, initial surgical exposure to the human brain appears to be a sophisticated factor itself which might determine the postoperative course and its length. The proposed model may still be improved by text preprocessing (e.g. lemmatization), adding convolutional neural networks layers, optimizing the architecture and training, testing alternative word embeddings (GloVe, Word2vec) [10].

5. Conclusion

Unstructured operative reports served as an informative source for prediction of in-hospital postoperative period in neurosurgery using deep learning. Our study demonstrated the usability of data stored in EHR as texts. *This research was supported by the Russian Foundation for Basic Research (grant 18-29-01052).*

References

[1] J. Senders, O. Arnaout, A. Karhade, H. Dasenbrock, W. Gormley, M. Broekman, T. Smith, Natural and Artificial Intelligence in Neurosurgery: A Systematic Review, *Neurosurgery*, **83**, (2018), 181–192.

[2] G. Danilov, K. Kotik, M. Shifrin, U. Strunina, T. Pronkina, A. Potapov, Prediction of Postoperative Hospital Stay with Deep Learning Based on 101 654 Operative Reports in Neurosurgery, *Stud. Health Technol. Inform.* **258** (2019), 125–129.

[3] M. A. Shifrin, E. E. Kalinina, E. D. Kalinin, A sustainability view on the EPR system of N.N. Burdenko Neurosurgical Institute, *Stud. Health Technol. Inform.* **129** (2007), 1214–1216.

[4] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, *arXiv Prepr. arXiv1607.01759*, 2016.

[5] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext. zip: Compressing text classification models, *arXiv Prepr. arXiv1612.03651*, 2016.

[6] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv Prepr. arXiv1412.3555*, 2014.

[7] J. Fürnkranz, Mean Absolute Error, *Encyclopedia of Machine Learning*, Springer US (2011), 652–652.

[8] W. E. Muhlestein, D. S. Akagi, J. M. Davies, L. B. Chambless, Predicting Inpatient Length of Stay After Brain Tumor Surgery: Developing Machine Learning Ensembles to Improve Predictive Performance, *Neurosurgery*, **85** (2019), 384-393.

[9] W. E. Muhlestein, D. S. Akagi, S. Chotai, и L. B. Chambless, The Impact of Race on Discharge Disposition and Length of Hospitalization After Craniotomy for Brain Tumor, *World Neurosurg.*, **104** (2017), 24–38.

[10] G. Danilov, M. Shifrin, U. Strunina, T. Pronkina, A. Potapov, An Information Extraction Algorithm for Detecting Adverse Events in Neurosurgery Using Documents Written in a Natural Rich-in- Morphology Language, *Stud. Health Technol. Inform.*, **262** (2019), 194–197.