ICT for Health Science Research
A. Shabo (Shvo) et al. (Eds.)

doi:10.3233/978-1-61499-959-1-125

Prediction of Postoperative Hospital Stay with Deep Learning Based on 101 654 Operative Reports in Neurosurgery

Gleb DANILOV[a,1], Konstantin KOTIK[b], 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*
[b] *Lomonosov Moscow State University, Moscow, Russian Federation*

Abstract. Electronic Health Records (EHRs) conceal a hidden knowledge that could be mined with data science tools. This is relevant for N.N. Burdenko Neurosurgery Center taking the advantage of a large EHRs archive collected for a period between 2000 and 2017. This study was aimed at testing the informativeness of neurosurgical operative reports for predicting the duration of postoperative stay in a hospital using deep learning techniques. The recurrent neuronal networks (GRU) were applied to the word-embedded texts in our experiments. The mean absolute error of prediction in 90% of cases was 2.8 days. These results demonstrate the potential utility of narrative medical texts as a substrate for decision support technologies in neurosurgery.

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

1. Introduction

The operative report is a document to describe the details of a surgery in a patient's medical records. Although operative reports are integrated with electronic health records (EHRs) as an essential component, they are often narrative with a huge variability of style and composition [1]. However, these unstructured medical records may potentially conceal a body of information and provide new knowledge when appropriately processed for the analysis.

The text is the most accepted form of information media in EHR, being poorly formalized though. Advanced text mining technologies enable exploration of medical records content, highlighting the relevant information for the generation and testing of research hypotheses. The data extracted from the original or preprocessed medical texts could be fed to machine learning algorithms for diagnostic and prognostic purposes [2]. We wondered if a description of neurosurgical procedures could bring a useful information to predict the duration of hospital stay following the operation. Inpatient length of stay may serve as an indirect estimate of patient disease severity and resource allocation [3]. Thus, the aim of our study was to find a model predicting the

[1] Corresponding Author: Gleb Danilov, E-mail: glebda@yandex.ru.

postoperative stay duration based on the operative reports using deep learning algorithms.

2. Methods

The EHRs of N.N. Burdenko Neurosurgery Center were searched to find all operative reports for neurosurgical procedures and related duration of in-hospital stay after surgery in the period between 2000 and 2017 [4]. The aim of the study was translated into a regression task for the machine learning. The target (outcome) variable was the length of stay at N.N. Burdenko Neurosurgery Center after the operation (number of days stored in EHR as integers). The texts of neurosurgical operative reports typed by neurosurgeons on computer keyboards were applied as an input to be fed into deep learning algorithms. All documents were preprocessed before piped in a learning procedure as follows: transformed to lower case, tokenized with a space separator (each document as an array of tokens) to form a dictionary, proceeded by word embeddings.

The tokens generated from the original texts were being matched with numeric vectors from Euclidean space R^d when d = 400. These vectors served as the inputs of subsequent models. Distributed word representation was obtained using the FastText library [5]–[7]. The idea of that approach was in learning to predict a word from its context or vice versa. However, we considered the basic vector representations of characters trigrams rather than of single words. We expanded each word as the sum of the vectors of its trigrams and then built the usual 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. As a result, we obtained vector representations which had been expected to work better with the rich-in-morphology languages, such as Russian.

A recurrent neuronal network (RNN) - gated recurrent unit (GRU) was applied to build a text-processing model [8], [9]. We used a logarithm of the hyperbolic cosine of the prediction error as a loss function. This function is approximately equal to the square of its argument for the small values, or to the difference between the modulus of the argument and natural logarithm of 2 for the large values of the argument. We applied the adaptive gradient method RmsProp 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]. The activation function used was Exponential Linear Unit [10]. The hyperparameters of the model were: batch size = 256 texts, dropout (for regularization) = 0.3. Mean Absolute Error (MAE) was considered as a measure of prediction precision calculated by the formula:

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

where n – was the number of operative reports, y_j – was the ground truth duration of the postoperative in-hospital stay, $\hat{y}_j$ – was the predicted duration of postoperative in-hospital stay.

To process the models a Keras framework (keras.io) was used. The model was trained on the first 500 vector-represented tokens corresponding to text protocols. In cases when the number of tokens exceeded 500, the first 500 tokens were fed into the model. Otherwise, when the number of tokens was less than 500, the sequence was padded with a fixed token so that the number of tokens in the sequence was still 500. The training sample was randomly generated from 75% of all operative reports. The testing sample constituted of remaining 25% of documents. Data preparation after the export from EHRs was done using R programming environment (version 3.5.0) in RStudio IDE for MacOS (version 1.1.453). The model was trained and evaluated within the Python (version 3.6) environment Jupyter Notebook for MacOS.

3. Results

A total of 104 506 neurosurgical operations were identified in the EHR system for a period between 2000 and 2017. In 101 681 (97,3%) cases, the full text of operative reports was applicable for further processing. The duration of postoperative hospital stay was available for 101 654 (97,3%) reports in the selected sample. The postoperative period ranged 0 - 1251 days per operative report (median = 8 [6;12]). The duration of stay after surgery did not exceed 119 days in 99% of cases, was limited up to 60 days in 97% of cases, was less than 42 days - in 95% of cases and did not lay beyond 23 days in 90% of observations. The distribution of the postoperative in-hospital stay matching with operative report texts is shown in Figure 1.

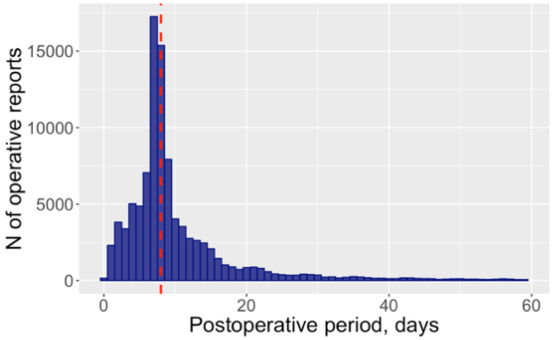

Figure 1. The distribution of the target variable (postoperative period per operative report), 60+ days excluded. Median value = 8 days (shown as a dashed red line).

Given such a variability in postoperative staying, we obtained 5 subsamples for machine learning experiments. The results of model training and testing for each subsample are summarized in Table 1.

Table 1. The results of RNN (bidirectional GRU) training and testing in 5 subsamples (AE – absolute error)

	Sample 1 (100%)	**Sample 2 (99%)**	**Sample 3 (97%)**	**Sample 4 (95%)**	**Sample 5 (90%)**
Sample size	104 654	100 644	98 641	96 657	91 871
Range of postoperative stay, days	0 - 1251	0 - 119	0 - 60	0 - 41	0 - 23
Testing sample size	26 123	25 861	25 341	24 845	23 527
MAE	7.78	5.43	4.54	3.82	2.77
Predictions with AE = 0 days	14.1%	18.07%	15.4%	17.5%	18.7%
Predictions with AE < 2 days	38.5%	44.0%	40.6%	43.9%	47.4%
Predictions with AE < 3 days	54.3%	59.2%	56.1%	58.2%	63.0%
Median AE, days	2.19	1.84	2.04	1.86	1.63
Maximum AE, days	1070.51	114.57	54.2	37.2	19.7

The best bidirectional GRU model showed the most impressive result in a Sample 5 with MAE of 2.8 days. In this sample it was possible to give an absolutely accurate prognosis in almost every 5th case and the precision did not exceed 1 day in almost 50% of cases. A K-fold validation (k = 10) improved the final MAE metric only by 0.03. The distribution density plots for absolute errors with MAEs obtained in 5 subsamples are shown in Figure 2. Our model obviously outperformed a naive estimator - median stays in a sample (e.g., MAE 7.78 days vs 10.53 days in Sample 1).

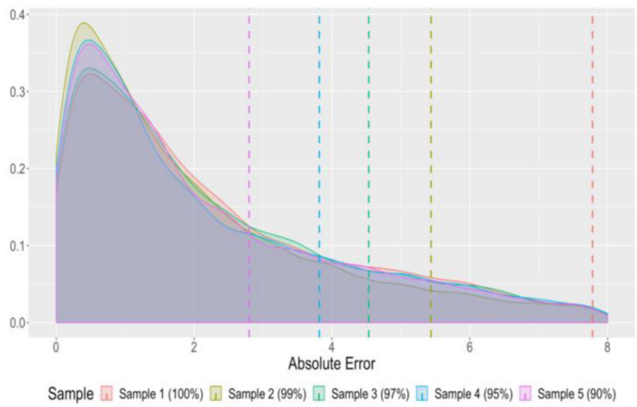

Figure 2. The density of absolute error distribution (the absolute difference between predicted and ground true values of postoperative stay, days) in Samples 1-5. The X-axis values are limited to 8. The dashed colored lines show the MAE for corresponding samples.

4. Discussion

The use of non-regression machine learning to model clinical outcomes in neurosurgery is not well-established [3], [12]. W. Muhlestein et al. obtained good results predicting the length of stay after brain tumor surgery using a set of machine learning methods trained on non-textual data [3], [12]. However, we could not identify any related publications on using unstructured medical text, natural language processing, and machine learning techniques to model the duration of the postoperative period in neurosurgery.

We found that the information concluded in neurosurgical operative reports was quite rich in its contents enabling the predictions coming from the "black box". However, the different results we obtained from a variety of subsamples could be reasonably explained. The marginal outliers in terms of postoperative days, which were cut off each time when subsetting 99%, 97%, 95% and 90% from the initial sample, appeared to be extremely rare and better explained by occasional factors not related to the initial surgery directly, such as severe complications and repeated operations. Thus, the clinical information linked to the prognosis directly in the majority of cases was not that relevant for the outliers. Alternatively, the percent of outlying values of a target variable could be not big enough to train the model appropriately on the whole dataset.

The GRU appeared to be appropriate model for the described machine learning task. It receives a sequence of numerical vectors representing words from a text protocol as the input, without violation of their order. It tends to preserve a knowledge from words while reading of the text. The GRU RNN architecture implements the idea of "text memorizing" while bidirectionally and independently "reading" the text from left to right and from right to left for each layer. The model itself could still be improved with its architecture by augmenting text preprocessing (e.g. words lemmatization), adding convolutional neural networks layers, optimizing the RNN architecture and training, test alternative word embedding techniques (GloVe, Word2vec). It would be also useful to conduct a "cleaner" experiment preselecting the operative reports separately for the first and repeated cases or merging them for one in-hospital episode. Although, not enough to build a decision support tool, the suggested GRU model demonstrated the usability of data stored as a text in EHRs. To understand what information (features) was extracted by neural networks from operative reports, various pre-and intraoperative data might be included in the analysis, such as the assessment of the intervention complexity, the duration of surgery, etc. The comparison of RNN prediction with neurosurgeon's expectation of a postoperative period duration based on the operative report reading would be of special interest.

5. Conclusion

The operative reports in neurosurgery appeared to be reasonably informative to predict the duration of postoperative hospital stay. Machine learning proved to be a promising technique to discover more hidden knowledge from unstructured medical data.

This research was supported by the Russian Foundation for Basic Research (grant 18-29-01052).

References

[1] G. B. Melton, N. E. Burkart, N. G. Frey, J. G. Chipman, D. A. Rothenberger, and S. M. Vickers, "Operative report teaching and synoptic operative reports: a national survey of surgical program directors.," *J. Am. Coll. Surg.*, vol. 218, no. 1, pp. 113–118, Jan. 2014.

[2] I.-H. Yoo and M. Song, "Biomedical ontologies and text mining for biomedicine and healthcare: A survey," *J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 109–136, 2008.

[3] W. E. Muhlestein, D. S. Akagi, J. M. Davies, and L. B. Chambless, "Predicting Inpatient Length of Stay After Brain Tumor Surgery: Developing Machine Learning Ensembles to Improve Predictive Performance," *Neurosurgery*, Aug. 2018.

[4] 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.*, vol. 129, no. Pt 2, pp. 1214–1216, 2007.

[5] "https://fasttext.cc/." [Online]. Available: https://fasttext.cc/.

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

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

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

[9] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv Prepr. arXiv1409.1259*, 2014.

[10] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural networks Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[11] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv Prepr. arXiv1511.07289*, 2015.

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