Improving the Prediction of Functional Outcome in Ischemic Stroke Patients

Miguel Monteiro*
INESC-ID
Lisbon, Portugal
mab.mtr@gmail.com

Ana Catarina Fonseca
IMM / School of Medicine, University of Lisbon
Lisbon, Portugal
acfonseca@medicina.ulisboa.pt

Ana Teresa Freitas
INESC-ID / IST, University of Lisbon
Lisbon, Portugal
atf@inesc-id.pt

Teresa Pinho e Melo
IMM / School of Medicine, University of Lisbon
Lisbon, Portugal
tmelo@medicina.ulisboa.pt

Alexandre P Francisco
INESC-ID / IST, University of Lisbon
Lisbon, Portugal
aplf@tecnico.pt

Jose M Ferro
IMM / School of Medicine, University of Lisbon
Lisbon, Portugal
jmferro@medicina.ulisboa.pt

Arlindo L. Oliveira
INESC-ID / IST, University of Lisbon
Lisbon, Portugal
aml@inesc-id.pt

ABSTRACT

Ischemic stroke is a leading cause of disability and death worldwide among adults. Despite advances in treatment, around one-third of surviving patients still live with long-term disability. The individual prognosis after stroke is extremely dependent on treatment decisions physicians take during the acute phase.

In the last five years several scores such as the ASTRAL, DRAGON and THRIVE scores have been proposed as tools to help physicians predict the patient functional outcome after three months of the initial stroke. These scores are rule based classifiers that use features available when the patient is admitted to the emergency room, and are selected or preselected by domain experts.

In this paper we apply machine learning techniques to the problem of predicting the functional outcome of ischemic stroke patients, three months after they were admitted for emergency treatment. We show that a pure machine learning approach achieves only a marginally superior Area Under the ROC Curve (AUC) (0.798) than that of the best score (0.782) when using features only available at admission. Furthermore, we show that by progressively adding features available at other points in time, we can significantly increase the AUC to above 0.90, a value well above the scores.

We conclude that the results obtained validate the use of the scores at the time of admission, but also point to the importance of using more features when possible, which require more advanced methods. We also conclude that machine learning techniques are instrumental tools that can be effectively used by physicians to improve the outcome prediction of ischemic stroke.

KEYWORDS

Ischemic Stroke, Machine Learning, L1 Regularization, AUC

ACM Reference format:
Miguel Monteiro, Ana Catarina Fonseca, Ana Teresa Freitas, Teresa Pinho e Melo, Alexandre P Francisco, Jose M Ferro, and Arlindo L. Oliveira. 2017. Improving the Prediction of Functional Outcome in Ischemic Stroke Patients. In *Proceedings of International Workshop on Data Mining in Bioinformatics, Halifax, Nova Scotia, Canada, August 2017 (BIOKDD'17),* 5 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND RELATED WORK

Ischemic stroke, which occurs due to an obstruction in blood vessels that supply blood to the brain, is a leading cause of disability and mortality worldwide among adults. Despite advances in treatment, around one-third of patients who survive will live with long term disability [19]. Patients or relatives frequently ask questions about the individual prognosis after stroke and if the patient will be functionally independent in his or her daily life. The answer to this question is not straightforward. Tools that could help determine, with high accuracy, what the functional prognosis of the patient will be after an ischemic stroke could be extremely useful.

Several efforts have been developed to create scores that can predict the patient's functional outcome defined as a discretized version of the modified Rankin Scale (mRS) [16], among them, the ASTRAL [15], DRAGON [18] and THRIVE [5] scores. These scores use statistical analysis to determine the most relevant covariates from a set of preselected features by domain experts. They are meant to be easily calculated by humans using data readily available when the patient is admitted to the emergency service, for this reason, the model's weights are discretized and the number

[*] To whom all correspondence should be sent.

BIOKDD'17, Halifax, Nova Scotia, Canada
© 2017 Copyright held by the owner/author(s). 978-x-xxxx-xxxx-x/YY/MM...$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

of covariates significantly reduced, which deteriorates their performance. From a machine learning perspective, these scores can be viewed as rule-based classifiers created by domain experts. All these scores have been externally validated and report an AUC in the range of 0.70 to 0.80 [4, 6, 7].

More complex studies, which take advantage of the longitudinal aspect of the data, and more powerful statistical tools also exist. For example one study [17] proposes a Markov chain like model which uses Bayesian inference to predict short-term stroke outcomes in a multi-classification environment (non-discretized mRS). This differs from our work in that our goal is to predict a long-term outcome and therefore we discretize the mRS since we are not interested in small variations between the classes considered in the aforementioned study. Other studies focus on different stroke outcome measures like the Barthel index [13]. In such a study [12] the authors take a similar approach to the scores that are used to predict the mRS.

As far as we know, there is no report of a pure machine learning approach to the problem of predicting the long-term functional outcome, nor a comparison between such an approach and the aforementioned scores [5, 15, 18]. In addition, it is also interesting to analyze how the prediction can improve when information collected at different stages of the treatment is used, and not just the information collected at admission time. Even though this prediction cannot be made at time of admission, like the scores, it can still be used to inform the physician throughout the course of the disease.

The goal of this paper is to use machine learning techniques to predict the functional outcome of a patient three months after the initial stroke, using information available at admission, and to compare the results with the predictions given by the scores. Furthermore, we analyze how the prediction improves as we add more features collected at different points in time after admission. We aim to show how machine learning techniques can be successfully applied to clinical data without losing interpretability of the models, a characteristic which is valuable for medical professionals.

2 METHODS

Our original dataset was comprised of 541 patients with acute stroke from the Safe Implementation of Treatments in Stroke (SITS) - Thrombolysis registry [20]. The cohort of patients came from the Hospital of Santa Maria, in Lisbon, Portugal, which is a tertiary university hospital. Even though different hospitals have different cohorts, all hospitals collect the same data for the SITS registry. This registry had baseline data collected at admission, follow-up data collected 2 hours, 24 hours, and 7 days after the initial stroke event, data collected on discharge, and the mRS recorded three months after the event. All of the patients on the registry were treated using Recombinant Tissue Plasminogen Activator (rtPA).

2.1 Data description and preprocessing

Data cleaning was performed by deleting features that contained only missing values and features which were meta-data (i.e. record number). In addition, we transformed categorical variables into dummy variables and converted variables that record times to time differences between variables (i.e. time of initial event and time of arrival at the hospital becomes time between event and arrival). In the end, the dataset had 147 features and 533 patients.

For data imputation, we used the median value for numerical variables (e.g. age) and the mode for binary variables (i.e. gender). All features were scaled to have zero mean and unit variance.

The features obtained could be divided into five sets, depending on the time at which they were collected:

1. Baseline: features collected at admission: the patient's demographic information, past history and risk factors; the time between stroke onset, arrival at the hospital and the start of treatment; the NIH Stroke Scale (NIHSS) [3] (discriminated by field, not just the final result); exam results; type of treatment;
2. 2 hours after admission: discriminated NIHSS and exam results;
3. 24 hours after admission: discriminated NIHSS, exam results, data relating to lesions detected during Computerised Tomography (CT) scan and/or Magnetic Resonance Imaging (MRI), the global outcome as reported by the physician;
4. 7 days after admission: discriminated NIHSS, the global outcome;
5. Discharge: discriminated NIHSS, the global outcome; the treatment the patient was discharged with; the results of cause investigation.

The target variable was the mRS three months after the event. To turn the problem into a binary classification problem, and to compare our results directly with the existing scores [5, 15, 18], we discretized the mRS into two classes according to:

- Good outcome: defined by $mRS \leq 2$
- Poor outcome: defined by $mRS > 2$

This particular discretization is of medical relevance, since it separates the patients who will be able to live a rather normal unassisted life, from the ones who will require significant assistance.

2.2 Experimental design

After preprocessing the data, we designed five experiments that aimed to assess with what precision we could predict the patient's mRS three months after admission, given information available at different points in time.

Table 1 presents the five experiments that where conducted as well as the total number of patients, class split percentages, and number of features used in each experiment.

Note that the total number of patients and class split percentage varies from experiment to experiment because at each time step we removed the patients who have died. We made this choice because, for these patients, the mRS at three months would be perfectly defined by this information alone and this would cause us to overestimate the performance of the methods.

For experiment 1, where only baseline information is available, we also calculated the ASTRAL, DRAGON and THRIVE scores to make possible a direct comparison.

2.3 The classifier

For the classifier, we choose to use logistic regression with L1 regularization, because this classifier achieves a good compromise between performance and interpretability. Since the features are

Table 1: Experiments.

Experiment Number	Feature sets used	# samples	Good outcome	Poor Outcome	# features
1	Baseline	533	51.3% (273)	48.7% (260)	48
2	Baseline, 2h	533	51.3% (273)	48.7% (260)	65
3	Baseline, 2h, 24h	532	51.4% (274)	48.6% (258)	99
4	Baseline, 2h, 24h, 7d	523	52.5% (275)	47.5% (248)	115
5	Baseline, 2h, 24h, 7d, discharge	507	54.6% (277)	45.3% (230)	147

scaled, it is possible to look at the absolute value and sign of the weights and determine the relative importance of the associated feature in the final outcome. It is important to note that L1 regularization performs automatic feature selection, by setting many of the weights to zero, and therefore, unlike the scores, it does not need expert knowledge to perform feature selection.

A brief mathematical description of the model is given below [14]. The logistic function is:

$$F(\boldsymbol{X}) = \frac{1}{1 + e^{-\boldsymbol{\beta}^T \boldsymbol{X}}}, \tag{1}$$

where:

$$\boldsymbol{X} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \tag{2}$$

In a binary classification problem, where y and $1 - y$ are the true class probabilities, and $\hat{y}$ and $1 - \hat{y}$ are the estimated class probabilities given by $\hat{y} = F(\boldsymbol{X})$, the cross entropy between the true and estimated probabilities is given by:

$$H(y, \hat{y}) = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}). \tag{3}$$

The cross entropy can be seen as a measure of the error between y and its estimate. The cost function of a logistic regression is then:

$$J(\beta) = -\frac{1}{m} \sum_{i=1}^{m} H(y^i, \hat{y}^i). \tag{4}$$

Adding the L1 regularization term:

$$J(\beta) = -\frac{1}{m} \sum_{i=1}^{m} H(y^i, \hat{y}^i) + \frac{1}{C} \|\beta\|_1, \tag{5}$$

where the L1 norm of the weights is multiplied by a reverse cost regularization constant C, a tunable hyper-parameter that controls how much weight the regularization term has on the overall cost function. The L1 norm of a vector is simply the sum of the absolute values of all its elements.

Since the $J(\beta)$ is a function of β, by minimizing the cost function we obtain the optimal set of weights which minimize the cross entropy between y and its estimates while keeping the size and the number of non-zero weights as small as possible.

To measure the performance of the model we used the AUC [1] since the accuracy is not a good measure of performance when the classes are imbalanced and the AUC is more informative [2]. We also evaluated the precision-recall trade-off curves.

Table 2: AUC for scores and Exp 1.

	Logistic	ASTRAL	DRAGON	THRIVE
Exp.1	**0.798**	0.782	0.764	0.750

To train and validate the model we used leave-one-out cross validation: we take a sample out of the dataset, train the model using all other samples and then validate the model on the sample that was excluded. The process is repeated for every sample and the predictions of each model are then aggregated to calculate the AUC of the method. Note that these metrics do not represent the performance of a single classifier but instead they approximate the performance of a classifier trained on all the available data and validated on unseen data.

To determine the best value for the reverse cost regularization constant C we performed a grid search over a set of reasonable values for C and chose the model which resulted in the highest AUC for the validation set.

For the sake of comparison, in the results section, we also present results for a decision tree and a Support Vector Machine (SVM), applied to the same dataset and using the same methodology (leave one out cross validation and a model search). For the SVM we searched over the cost parameter and the linear and radial kernels. For the decision tree we searched over the maximum number of features considered per split, the minimum number of samples per leaf, the maximum depth of the tree and the minimum number of samples per split.

3 RESULTS

Table 2 and Figure 1 present the AUC and Receiver Operating Characteristic (ROC) curves for the ASTRAL, DRAGON and THRIVE scores, as well as the logistic regression when applied to the baseline data (Exp. 1).

From Table 2 and Figure 1 we can see that, even though the logistic has the highest AUC 0.798, the ASTRAL score is very close with an AUC of 0.782.

Table 3 presents the AUC for the five experiments and three different classifiers. Figure 2 and Figure 3 show the ROC and precision-recall curves for the same experiments, but only for logistic regression.

From Table 3 we can see that the SVM performs better for the first two experiments but the logistic regression gains the lead after experiment 2. As expected, decision trees performed poorly, despite their interpretability, and SVMs performed well but have low interpretability.

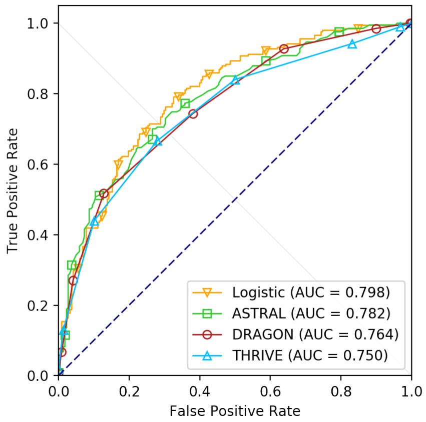

Figure 1: ROC curves for scores and Exp 1.

Table 3: AUC for different classifiers and experiments.

	Logistic	SVM	Tree
Exp. 1	0.798	**0.806**	0.687
Exp. 2	0.834	**0.843**	0.725
Exp. 3	**0.899**	0.885	0.846
Exp. 4	**0.925**	0.916	0.862
Exp. 5	**0.922**	0.912	0.831

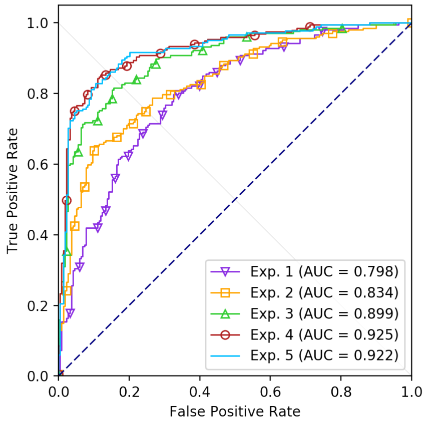

Figure 2: ROC curves for logistic regression.

From Figure 2 it is possible to see that, up to experiment 4, the more data we add the greater the AUC. There are two big jumps in performance: when we add the 2 hours feature set and then again when we add the 24 hours feature set. We note that for experiments 4 and 5 the AUC is above 0.90.

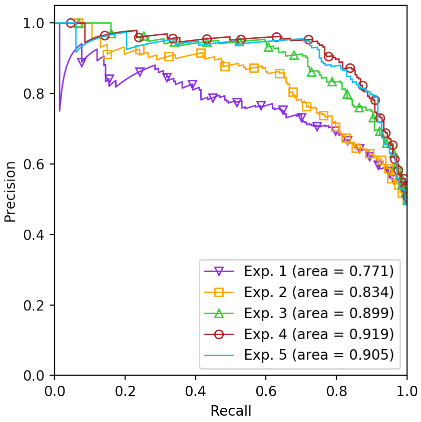

Figure 3: Precision-recall curves for logistic regression.

Table 4: Feature importance (Exp. 4)

rank	Exp. 3
1	nih5A_7d (0.92)
2	nih5B_24h (0.54)
3	nih2_7d (0.34)
4	globalOutcome_7d (-0.30)
5	Age (0.26)
⋮	⋮
20	nih11_24h (0.04)

From Figure 3 we see that the precision-recall curves follow the same performance order as their ROC counterparts. For experiments 4 and 5 the area under the curve is above 0.90 and the precision only starts to drop significantly when the recall is already very high (~0.80).

Since all the features are scaled to zero mean and unit variance we can look at the weights of a model to assess the relative importance of a feature in the final prediction. The top five weights sorted by absolute value for the model with the best result (Exp. 4) are shown in Table 4.

From Table 4 we can see that the NIHSS fields 5A and 5B which relate to left and right arm motor strength have a large importance in the prediction. A detailed discussion of this result is presented in Section 4.

4 DISCUSSION

The results of Table 2 and Figure 1 validate the use of scores when only data collected at admission is available. However, by comparing the results of the scores with the results of experiments 2-5 it becomes clear that the use of these scores is not the best option when more data is available. When more information is available, logistic regression is able to achieve an area under the curve over 0.90 for both the ROC and the precision-recall curves, which is a good result given the significant uncertainties inherent to the problem. This information can be effectively used by the physician to make medical decisions, regarding prognosis, and to better inform the patient and the family about the predictable functional outcome. In addition, we believe that by adding more data to the classifier, we could improve its performance relative to the scores even when only admission data is available.

By looking at Table 4, we see that the features with the highest weight associated with the outcome were motor arm strength (nih5A and nih5B). There is biological plausibility for the selection of these variables, since arm strength is essential to the performance of most daily activities that are evaluated by the mRS. Neglect (nih11) is related to a higher cortical function, and is mainly due to injuries to the right cerebral hemisphere. It may behave as a confounding variable, a factor that explains the preponderance that was found for the left arm (nhi5A). Neglect refers to a tendency to exhibit decreased surveillance of the left side of personal and extra personal space. In neglect, stimuli from specific modalities (motor, visual, auditory or sensation) are not properly processed . Evaluation of neglect is more technically demanding that the evaluation of other neurological deficits [9]. It is possible that some evaluators underreported the presence of this neurological deficit, which is frequently associated to left paresis, explaining why neglect does not stand as an independent predictor of the outcome by itself in this study (rank 20). Neglect has been previously associated, in observational studies, to depressed activities of daily living and impaired functional outcome and worse outcome rehabilitation [10, 11]. Age at the time of development of a cerebral lesion has also been previously described as an important determinant of outcome. Younger patients tend to recover more easily and completely than older patients [8].

5 CONCLUSION AND FUTURE WORK

We conclude that machine learning techniques can be used to predict the functional outcome of ischemic stroke three months after the event, with an AUC ranging from 0.798 to 0.925, depending on the point in time at which the prediction is made. Furthermore, we have validated the use of scores when only data at admission is available and have shown that logistic regression with L1 regularization is an interpretable model from which experts can derive new knowledge.

In the future we wish to improve our prediction by using more and richer records, both by using more cohorts from the SITS database and by using our own database which we are currently developing with more complex data. Moreover, we aim to incorporate the use of image and genetic information and to take advantage of the longitudinal aspect of the data.

ACKNOWLEDGMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and by project PRECISE - accelerating progress toward the new era of precision medicine (PAS application C443089117-00042059).

REFERENCES

[1] Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology 12*, 4 (1975), 387–415.

[2] Bradley, A. A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms, 1997.

[3] Brott, T., Adams, H. P., Olinger, C. P., Marler, J. R., Barsan, W. G., Biller, J., Spilker, J., Holleran, R., Eberle, R., and Hertzberg, V. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke 20*, 7 (1989), 864–870.

[4] Cooray, C., Mazya, M., Bottai, M., Dorado, L., Skoda, O., Toni, D., Ford, G. A., Wahlgren, N., and Ahmed, N. External Validation of the ASTRAL and DRAGON Scores for Prediction of Functional Outcome in Stroke. *Stroke 47*, 6 (2016), 1493–1499.

[5] Flint, A. C., Cullen, S. P., Faigeles, B. S., and Rao, V. A. Predicting long-term outcome after endovascular stroke treatment: The totaled health risks in vascular events score. *American Journal of Neuroradiology 31*, 7 (2010), 1192–1196.

[6] Flint, A. C., Faigeles, B. S., Cullen, S. P., Kamel, H., Rao, V. A., Gupta, R., Smith, W. S., Bath, P. M., and Donnan, G. A. Thrive score predicts ischemic stroke outcomes and thrombolytic hemorrhage risk in vista. *Stroke 44*, 12 (2013), 3365–3369.

[7] Gaucher, P., and Hildncr, T. Totaled Health Risks in Vascular Events Score Predicts Clinical Outcome and Symptomatic Intracranial Hemorrhage in Chinese Patients After Thrombolysis. *Tanaka et al. 18*, 6 (2015), 11.

[8] Harvey, R. L. Predictors of Functional Outcome Following Stroke, 2015.

[9] Jehkonen, M., Ahonen, J.-P., Dastidar, P., Koivisto, A.-M., Laippala, P., and Vilkki, J. How to detect visual neglect in acute stroke. *The Lancet 351*, 9104 (may 1998), 727–728.

[10] Jehkonen, M., Ahonen, J.-P., Dastidar, P., Koivisto, A.-M., Laippala, P., Vilkki, J., and Molnar, G. Visual neglect as a predictor of functional outcome one year after stroke. *Acta Neurologica Scandinavica 101*, 3 (2000), 195–201.

[11] Jehkonen, M., Laihosalo, M., and Kettunen, J. E. Impact of neglect on functional outcome after stroke - A review of methodological issues and recent research findings. *Restorative Neurology and Neuroscience 24*, 4-6 (2006), 209–215.

[12] Lewis, S. C., Sandercock, P. A., and Dennis, M. S. Predicting outcome in hyper-acute stroke: validation of a prognostic model in the Third International Stroke Trial (IST3). *J Neurol Neurosurg Psychiatry 79*, 4 (2008), 397–400.

[13] Mahoney, F. I., and Barthel, D. W. Functional evaluation: The barthel index. *Maryland state medical journal 14* (1965), 61–65.

[14] Murphy, K. P. *Machine Learning: A Probabilistic Perspective.* Adaptive computation and machine learning. MIT Press, 2012.

[15] Ntaios, G., Faouzi, M., Ferrari, J Lang, W., Vemmos, K., and Michel, P. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. *Neurology 78*, 2 (2012), 1916–22.

[16] RANKIN, J. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. *Scottish medical journal 2*, 5 (may 1957), 200–215.

[17] Sengupta, A., Rajan, V., Bhattacharya, S., and Sarma, G. R. K. A statistical model for stroke outcome prediction and treatment planning. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2016), pp. 2516–2519.

[18] Strbian, D., Meretoja, A., Ahlhelm, F. J., Pitkäniemi, J., Lyrer, P., Kaste, M., Engelter, S., and Tatlisumak, T. Predicting outcome of IV thrombolysis - Treated ischemic stroke patients: The DRAGON score. *Neurology 78*, 6 (2012), 427–432.

[19] Truelsen, T., Piechowski-Jóźwiak, B., Bonita, R., Mathers, C., Bogousslavsky, J., and Boysen, G. Stroke incidence and prevalence in Europe: a review of available data. *European journal of neurology : the official journal of the European Federation of Neurological Societies 13*, 6 (2006), 581–98.

[20] Wahlgren, N., Ahmed, N., Dávalos, A., Ford, G. A., Grond, M., Hacke, W., Hennerici, M. G., Kaste, M., Kuelkens, S., Larrue, V., Lees, K. R., Roine, R. O., Soinne, L., Toni, D., and Vanhooren, G. Thrombolysis with alteplase for acute ischaemic stroke in the Safe Implementation of Thrombolysis in Stroke-Monitoring Study (SITS-MOST): an observational study. *Lancet 369*, 9558 (2007), 275–282.