Comparison of Bio-inspired Algorithms Applied to the Hospital Mortality Risk Stratification

Jesús Silva, Yaneth Herazo-Beltrán, Freddy Marín-González, Noel Varela, Omar Bonerge Pineda Lezama, Pablo Palencia, and Carlos Vargas Mercado

Abstract The construction of patient classification (or risk adjustment) systems allows comparison of the effectiveness and quality of hospitals and hospital services, providing useful information for management decision making and management of hospitals. Risk adjustment systems to stratify patients' severity in a clinical outcome are generally constructed from care variables and using statistical techniques based on logistic regression (RL). The objective of this investigation is to compare the hospital mortality prediction capacity of an artificial neural network (RNA) with other methods already known.

Keywords Hospital mortality · Risk stratification · Intensive care unit · Artificial neural networks · Bootstrap

J. Silva (✉)
Universidad Peruana de Ciencias Aplicadas, Lima, Peru
e-mail: jesussilvaUPC@gmail.com

Y. Herazo-Beltrán
Universidad Simón Bolívar, Barranquilla, Colombia
e-mail: aherazo4@unisimonbolivar.edu.co

F. Marín-González · N. Varela
Universidad de La Costa (CUC), Barranquilla, Colombia
e-mail: fmarin1@cuc.edu.co

N. Varela
e-mail: nvarela2@cuc.edu.co

O. B. P. Lezama
Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

P. Palencia
Corporación Universitaria Minute de Dios. UNIMINUTO, Barranquilla, Colombia
e-mail: pablo.palencia@uniminuto.edu.co

C. V. Mercado
Corporación Universitaria Latinoamericana, Barranquilla, Colombia
e-mail: cvargas@ul.edu.co

Á. Rocha et al. (eds.), *Developments and Advances in Defense and Security*,
Smart Innovation, Systems and Technologies 181,
https://doi.org/10.1007/978-981-15-4875-8_16

1 Introduction

Artificial neural networks (RRs) are calculation systems that resemble biological neural networks when using interconnected nodes (neurons). These nodes receive the information, perform operations on the data, and transmit their results to other nodes. The procedure is to train the RFs to learn complex patterns of relationships between the predictor and outcome variables, and they are able to face new data by giving the expected answers [1].

They are defined as nonlinear, flexible, and highly generalized systems. These properties have led to their dissemination in all scientific fields and for their equivalence or superiority to some statistical techniques to be demonstrated. Interest in the application of RNAs in medicine over the past 10 years has only increased, as reflected in the progressively increasing number of publications that include this methodology. The areas they have been occupying are image recognition, wave analysis, pharmacology procedures, epidemiology, prediction of results, and diagnostic processes [2, 3].

The use of RFs for risk stratification offers as an advantage a possible increase in predictive power (accuracy), which has been assessed by 5–10%, as they do not work with the rigid limitations of statistical models. Compared to RL techniques, RRs take into account nonlinear relationships, automatically, without the need to follow a particular model, and the possible interdependence of input variables [4].

Intensive care unit (ICU) is commonly used for probability of death calculation systems (as a criterion of severity for the sick), and one of the most common systems is acute physiology and Chronic Health Evaluation II (APACHE II) built with RL technique. The objectives of our work are to demonstrate the usefulness of the neural network-based methodology for risk stratification, applying it to the calculation of probability of hospital mortality, using the APACHE II system variables in ICU (as a sample of study). A logistic regression model is used for reference.

2 Methodology

2.1 Subject

The study was carried out in a multipurpose ICU in Colombia. Patients have been studied over a 5-year period (2010–2015). Coronary patients, heart surgery subjects, and burns are not included. Patients under the age of 16 have been excluded, those who have moved and those who have stayed less than 24 h admitted. Only the first admission has been taken into account in patients who re-enter. The demographic, evolution, and severity variables necessary for the calculation of the APACHE II system have been collected prospectively by a trained team. The conduct of the study was approved by the hospital's ethics committee ensuring at all times the anonymity of patients [1].

2.2 Variables

14 physiological variables are used (extended from 12 to 14, since to assess oxygenation we use PaO_2, plus FiO_2 and $PaCO_{2),}$ age and two variables to determine score according to chronic disease (chronic disease and urgent/programmed) that complete 17 input variables. The output variable is hospital mortality. The calculation of the probability of death based on APACHE II is done as standard, converting the APACHE II score and applying the logistic formula with the coefficients published in the original article of [5]; adjustment by diagnostic groups as it would motivate the addition of more than 40 variables. The 2325 patients who met the inclusion criteria were randomly assigned, 75% to the development group, and the remaining 25% to the validation group.

2.3 Bootstrap Resampling of the Development Group

As the number of cases available for the development of the model is limited, there is a risk that they will have poor representativeness of the population; therefore, we must use techniques that optimize the available data to achieve a good generalization. As a solution, we apply resampling techniques (bootstrap) that have proven useful for this purpose. In our case, to achieve sufficient accuracy, this resampling must be repeated at least 300 times [6].

2.4 Logistic Regression Model

A multiple logistic regression model is used with the addition of all variables (full model). The calculation will be made on the 300 bootstrap samples in the development group. The resulting 300 models (their coefficients) will be used to calculate the probabilities in the original development set. With these 300 probabilities we calculate the average probability, expressed as P-RL-D, and its standard error [7].

2.5 Artificial Neural Network Model

The type of network used is a multilayer perceptron trained with backpropagation algorithm and sigmoid activation function. We use a 3-layer model (input, hide, and output). The selection of the optimal architecture and parameters is based on an empirical procedure and cross-validation. The development set is divided (50% to ensure representativeness) into one training set and a verification set. This split is done randomly and is repeated 10 times to compare the results on these 10 occasions [8].

Supervised training involves the repeated presentation of the training set to the network; in each iteration, an adjustment of the weights is made to minimize the network error function. Weights are the internal values of the network that resemble the synaptic forces of biological models. The evaluated cost function or error function (both in the training and verification set) is the root of the mean quadratic error (ECM) between the predictions and the actual values. Nodes are added or removed from the hidden layer until the ECM (in the verification set) is minimized, which also determines the time to stop training. Other parameters that are modified during the training process (learning coefficient, moment, etc.) are adjusted to achieve this optimization [9].

2.6 Artificial Neural Network Training

The training conditions set out in the previous point will be used to train 300 networks with the data of the 300 bootstrap resamples of the development set. When these networks are faced with data from the original development set, they determine 300 probabilities and their mean is called P-RNA-D [10].

2.7 Model Validation

Both the 300 logistic regression models and the 300 trained networks must face data from the validation set. The calculated average probabilities will be identified as P-RL-V and P-RNA-V, respectively [11].

2.8 Comparison of Models

To compare the different models, their properties of discrimination will be measured by means of the area under the ROC curve (ABC) and the calibration with the Hosmer-Lemeshow C22 goodness-of-fit test, the construction of the calibration curves, and the calculation of the reasons for standardized mortality (MeR, which is the ratio between the number of deaths observed and the number of deaths expected according to the prediction model) with their confidence intervals. We use the Bland–Altman test to evaluate the concordance between the probabilities obtained by each model. We define extreme case when the patient reaches a probability difference between the RL model and the RNA model with absolute value equal to or greater than 0.2. Statistical calculations were performed with the SPSS 14.0 program. The program used for networking is Qnet 97 (Vesta Services Inc.) [12].

3 Results

Artificial neural network model The selection method led us to an optimal architecture with 9 nodes in the hidden layer and fully interconnected nodes. The starting point of the training was set at 1500 iterations. The parameters of the network training process were a learning coefficient of 0.01 and the time of 0.3.

3.1 *Comparison of Models*

Table 1 shows the results of the comparison between the various models. Good results are appreciated, both in discrimination and calibration, of the APACHE II system (in the development and validation set). Significantly higher values are observed in the ABC ROC RNA compared to the APACHE II model. The RNA, compared to the RL model, shows better values (which are maintained in the validation group) in both discrimination and calibration, although these values do not reach a significant difference. Figure 1 shows the ROC curves and calibration curves for the validation group.

P-AP-II: probability of the APACHE II model (confidence interval [IC] calculated according to Hanley and McNeil's work); P-RL: average probability of 200 bootstrap logistic regression models; P-RNA: average probability of 200 bootstrap models of artificial neural network; development: n × 800 patients; validation: n × 346; ABC: area under the ROC curve (95% CI); HL-C: Hosmer-Lemeshow C test with 8 degrees of freedom for the development group and 10 for the validation group ($p > 0.05$ determines a correct test); RME: standardized mortality ratio (95% CI).

Table 1 Comparison of results of the different models of probability of death calculation

	P-AP-II	P-RL	P-RNA
ABC (IC of the 95%)			
Development	0.82 (0.77–0.83)	0.81 (0.79–0.85)	0.87 (0.84–0.91)
Validation	0.80 (0.74–0.83)	0.82 (0.75–0.83)	0.81 (0.78–0.83)
HL-C (p)			
Development	17.44 (0.040)	12.23 (0.132)	6.68 (0.468)
Validation	12.38 (0.295)	29.58 (0.001)	11.32 (0.325)
RME (IC del 95%)			
Development	1.08 (0.95–1.21)	1.04 (0.96–1.20)	1.02 (0.98–1.10)
Validation	1.21 (1.04–1.30)	1.17 (1.01–1.23)	1.14 (0.98–1.19)

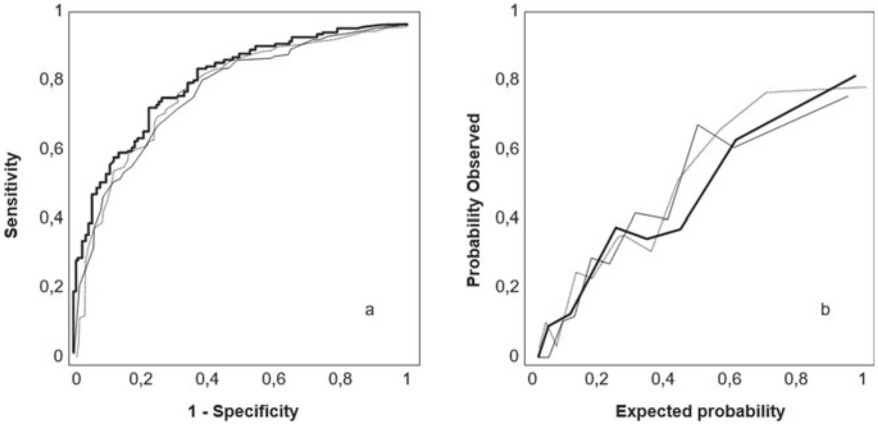

Fig. 1 ROC curves and calibration curves in the validation group. **a** ROC curves. **b** Calibration curves. Single line: Probability APACHE II; dashed line: Average probability of the bootstrap logistic regression model; thick line: Average probability of the bootstrap artificial neural network model

3.2 Analysis by Diagnostic Groups

In traumatic patients (28%), no model achieves sufficient accuracy (the APACHE II system is shown as the best model, but with ABC ROC less than 0.77 and poor calibration in both the development and validation assembly). In respiratory patients (26%), the best results are achieved with neural networks; and within this group, patients with chronic obstructive pulmonary disease (COPD) (14%) they are the worst performing with APACHE II and with RL maintaining acceptable RNA properties (ABC s 0.86 [0.77–0.92]) [4].

3.3 Comparison of Probabilities Between Models

The Bland–Altman technique (Fig. 2) shows us the lack of consistency in the probability assignment between the RL and RNA models (results shown in the validation set). The highest match tends to occur at the low values of the probability range and is lost by exceeding the calculated 36% chance of death. 95 patients (7%) were identified from the total study group as extreme cases. It is difficult to analyze the interrelationship between variables, but we find that in the subgroup of these patients where the probability assigned by RL is clearly higher than that calculated by RNA (42 patients), most (43 patients) have neurological alterations, and we appreciate that the Glasgow variable becomes more important in the RL model while maintaining similar values for the rest of the variables.

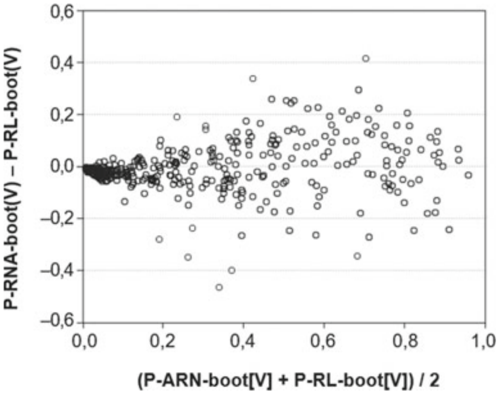

Fig. 2 Bland–Altman test (validation group) between the probabilities calculated by the logistic regression method versus the probabilities calculated by artificial neural network model. The dotted lines differentiate (above or below) to 2 standard deviations—the mean of the difference between the probabilities; P-RNA-V: Average probability of 300 bootstrap trained networks; P-RLV: Average probability of 300 bootstrap logistic regression models

4 Conclusions

The first analysis of our results is aimed at the good results found with the APACHE II model, since this fact did not coincide with other results that we had obtained with smaller series that analyzed less years in our database. Our ICU is characterized by fewer surgical patients scheduled and more mortality than the series that originally served to make the APACHE II system. We used this argument to justify, above all, deviations in calibration, which is almost optimal now that we can analyze more patients. This supports the concept of the large dependence of the sample size on any risk stratification analysis.

The high mortality found in our series is also conditioned by not including coronary patients. We were unable to include them due to follow-up care issues, given the characteristics of this group of patients' care in our hospital. As a global result, in our work we find better results with the RNA-based methodology, although they do not become significant. This result is similar to that achieved with other series. In a review conducted by [13], which analyzed 28 studies in cancer patients, it concludes that the networks are equivalent to or slightly higher than RL, as they do not have to rely on rigid requirements for variable or model independence Linear.

We also see that despite using resampling and cross-validation techniques we have, in our series, some "overlearning" problem: The network learns the patterns of the training set very precisely, but loses in the ability to generalization when faced with new data from the validation set. Working with networks, the conditions necessary to achieve a good generalization focus on three aspects: (a) that the information collected in the data is sufficient (this affects the size of the series and the quality in the collection of data); (b) that the "function" learned by the network is smooth (small changes in input variables will cause small changes in the output variable), and (c) that the size of the training set is sufficient and representative of the total data. The required size is determined by the number of network parameters, and 5 records are required per estimated parameter. In our example with 17 input variables, 9 hidden nodes, and one output node (which are 162 parameters), 800 cases are sufficient [14].

The algorithm we propose meets these conditions when working with limited series and can be applied in another type of population or health problem. There are other procedures based on different resampling and learning techniques that have been applied in other populations. Obtaining similar results, in the properties of discrimination and calibration with statistical and neural prediction models, has led some authors to claim that the relationship between variables is independent and virtually linear [15].

We provide the view that being able to obtain different individual probabilities implies that the relationship between variables is different when applying RL or RNA. It is true that the interpretation of this interrelationship is difficult (black box concept of RNA) [16], but we can study patients who define themselves as extreme cases (patients with neurological problems in our series), or compare the different behavior according to diagnostic groups (e.g., our different outcomes in traumatic and COPD patients).

References

1. Sargent, D.J.: Comparison of artificial neural networks with other statistical approaches results from medical data sets. Cancer **91**, 1636–1642 (2001)
2. Bifet, A., De Morales, G. F: Big data stream learning with Samoa. Recuperado de (2014). https://www.researchgate.net/publication/282303881_Big_data_stream_learning_with_SAMOA
3. Clermont, G., Angus, D.C., DiRusso, S.M., Griffin, M., Linde-Zwirble, W.T.: Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models. Crit. Care Med. **29**, 291–296 (2001)
4. Wong, L.S.S., Young, J.D.: A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural network. Anaesthesia **54**, 1048–1054 (1999)
5. Bravo, M., Alvarado, M.: Similarity measures for substituting web services. Int. J. Web Serv. Res. **7**(3), 1–29 (2010)
6. Chen, L., Zhang, Y., Song, Z.L., Miao, Z.: Automatic web services classification based on rough set theory. J. Cental South Univ. **20**, 2708–2714 (2013)
7. Viloria, A., Lezama, O. B. P: Improvements for determining the number of clusters in k-means for innovation databases in SMEs. ANT/EDI40, pp 1201–1206 (2019)
8. Viloria, A., Lis-Gutiérrez J. P., Gaitán-Angulo, M., Godoy, A. R. M., Moreno, G. C., Kamatkar, S. J.: Methodology for the design of a student pattern recognition tool to facilitate the teaching—Learning process through knowledge data discovery (big data). In: Tan, Y., Shi, Y., Tang, Q. (eds.) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham (2018)
9. Zhu, J., Fang, X. et al.: IBM cloud computing powering a smarter planet. In: Libro Cloud Computing, vol. 599.51, pp 621– 625 (2009)
10. Mohanty, R., Ravi, V., Patra, M.R.: Web-services classification using intelligent techniques. Expert Syst. Appl. **37**(7), 5484–5490 (2010)
11. Thames, L., Schaefer, D.: Software defined cloud manufacturing for industry 4.0. Procedía CIRP **52**, 12–17 (2016)