PREDICTION OF MORTALITY IN AN INDIAN INTENSIVE CARE UNIT: COMPARISON BETWEEN APACHE II AND ARTIFICIAL NEURAL NETWORKS

Ashish Nimgaonkar*, Sudarshan**, Dilip Karnad***

From *School of Biomedical Engineering and **Department of Computer Science and Engineering, Indian Institute of Technology, Powai, Mumbai and ***Department of Medicine, KEM Hospital, Parel, Mumbai 400012

ABSTRACT

Background - Rule-based decisions are not always accurate. Statistical techniques like multivariate regression have therefore been used to develop outcome prediction scores like APACHE, SAPS and MPM. However, these statistical methods are constrained by the assumption of predefined mathematical relationship between independent variables and outcome. Experienced doctors make better decisions than newcomers. Computer-based artificial intelligence techniques like artificial neural networks (ANN) can be "trained" to develop their own models to predict outcome based on large amounts of clinical data. The commonest learning mechanism in ANN is the backpropagation algorithm, wherein the system predicts the outcome for each patient based on past experience and compares this with actual outcome. In this algorithm, an error is propagated backwards to improve the prediction accuracy.

Objectives - To compare APACHE II (Acute Physiological and Chronic Health Evaluation) system with backpropagation ANN outcome prediction.

Design - Retrospective cohort study.

Setting - Medical intensive care unit, King Edward Memorial (KEM) Hospital, Mumbai.

Subjects - 2962 consecutive admissions between 1996 and 1999.

Methods - 22 clinical and laboratory variables were used to obtain the Day 1 APACHE II score and predict survival. The same 22 variables were also used as input for the ANN model (ANN_{22}). Data from the first 1962 patients were used to train the network and the next 1000 were used for validation. Calibration (degree of correspondence between estimated and observed mortality over each decile of risk) was determined by the Hosmer-Lemeshow H statistic; higher values indicate poorer goodness-of-fit. Discrimination (ability to distinguish between survivors and nonsurvivors) was estimated by the area under the ROC curve; values closer to 1 indicate better discrimination.

Results - 337 of the 1000 patients in the validation group died. APACHE II predicted 246 deaths while ANN-22 predicted 336 deaths. Calibration was better with ANN_{22} (H=22.4) than with APACHE II (H=123.5) and so was discrimination (area under ROC curve = 0.87 v/s 0.77 with APACHE II, p=0.002). Analysis of information gain (entropy) contributed by each of the 22 variables revealed that the ANN could reliably predict outcome using only 15 variables without loss of accuracy. The new model, ANN_{15}, had calibration (H=27.7) and discrimination (area under ROC 0.88) comparable to ANN_{22} (p=0.87) and was superior to the APACHE II model (P<0.001).

Conclusions - Both ANN models built from the Indian cohort of ICU patients outperformed the APCHE II system in Indian patients. The neural network could accurately predict outcome using only 15 variables.

INTRODUCTION

Rule-based decisions are not always accurate, especially when the interplay of many variables determines the ultimate outcome. Statistical techniques like multivariate regression have therefore been used to develop outcome prediction scores like APACHE, SAPS and MPM where age, pre-existing chronic diseases, clinical parameters, physiological derangements, surgical status and acute problems requiring ICU admission have been used to predict survival or death [1]. However, these statistical methods are constrained by the assumption of predefined mathematical relationship between independent variables and outcome [2]. Experienced doctors make better decisions than newcomers. Computer-based artificial intelligence techniques attempt to simulate the decision making processes such as learning and memory that are employed by the human brain [3]. One such technique is the artificial neural network (ANN). Computers with appropriate software, can be "trained" to develop their own models to predict outcome based on large amounts of clinical data. The commonest learning mechanism in ANN is the backpropagation algorithm, wherein the system predicts the outcome for each patient based on past experience (memory) and compares this with actual outcome. In this algorithm, an error is propagated backwards to improve the prediction accuracy [2,3].

It would be reasonable to expect that artificial intelligence methods may be able to predict ICU outcomes better than statistically derived equations. We therefore attempted to compare the predictive accuracy of ANNs with the Acute Physiological and Chronic Health Evaluation II (APACHE II) scoring system. The APACHE II was selected as it is still being widely used and considered the benchmark scoring system [4]. It has been studied in various countries other than the US, including UK, France, Germany, Saudi Arabia, Japan, New Zealand, Brazil and India [4,5].

MATERIALS AND METHODS

All patients aged 12 years and above, admitted to the 17 bedded Intensive Care Unit of King Edward Memorial, Mumbai between January 1996 and May 1998, were studied. The KEM Hospital is 1800 bedded Municipal hospital and tertiary referral center. The raw data used to obtain the APACHE II score were prospectively collected in all patients admitted to the ICU. The values recorded were the most abnormal physiological values during the first 24 hrs of ICU admission. The Hospital outcome (discharged or dead) was also recorded. The probability of hospital mortality, p, was derived from the APACHE II equation [6]:

$$\ln(p/(1-p)) = -3.517 + (0.146 \times \text{Apache score}) + (0.603 \times \text{post-emergency operation status}) + \text{disease category coefficient.}$$

ANNs have the ability to "learn" mathematical relationships between a series of input (independent predictor variables) and the corresponding output (dependent, outcome) variables. This is achieved by "training" the network with a training data set consisting of predictor variables and the known or associated outcomes. Networks are programmed to adjust their internal weights based on the mathematical relationships identified between the inputs and the outputs in a data set. Once a network has been trained, it can be used for pattern recognition or classification tasks in a separate test data set.

A variety of learning algorithms exist to determine the optimal network weights, but the most commonly and widely used algorithm is the backpropagation algorithm. The algorithm may be characterized as follows. A training case or pattern (i.e. an observation) is selected from the training set (randomly or sequentially) and then the values of the predictor variables are used as inputs to the network to determine the predicted output of the network. This predicted output is then compared to the known or desired output (e.g., 0=live, 1=die). If there is a difference between the predicted and the known output, an error signal is calculated and back-propagated through the neural network. This signal is a mathematical factor that adjusts the value of each weight in the neural network to reduce the difference between the predicted and the known output. Thus, the next time the network encounters a training case with similar input values, its prediction will be more accurate. All the cases in the training set are continually presented to the network until the overall error has been minimized. One pass through all of the training cases is considered an epoch of training. The overall duration of training is often expressed in terms of number of epochs required to reach an error minimum. The main power of neural networks lies within the hidden nodes. They are hidden from the input and output of the neural network and act as data feature detectors. It is here that the network is able to learn implicitly any arbitrarily complex non-linear relationship between the dependent and the independent variables.

In addition to learning to predict outcome, the network also determines the entropy (information gain) of the input variables. The entropy for a Boolean classification of a collection S, containing positive and negative examples of some target concept, can be given as

$$\boxed{Entropy\,(\mathbf{S}) = -\,p^{+}\log_2 p^{+} - p^{-}\log_2 p^{-}}$$

where p^{+} is the proportion of positive examples in the collection **S** and p^{-} is the proportion of negative examples in **S**.

In this study, 1000 cases were randomly selected and kept aside to serve as the test set (patients in whom the ANN would be used to see if it could correctly predict outcome). The remaining 1962 cases were taken as the training set. The neural network model was developed based on these 1962 cases. Initially all the 22 variables (attributes) that are used to predict mortality in the APACHE II model were taken to develop the network (ANN $_{22}$). Then of the 22 variables, only 15 variables with highest entropy (information gain) were selected for model building (ANN $_{15}$).

All three models, namely the APACHE II, ANN with 22 input variables (**ANN$_{22}$**) and ANN with 15 input variables (**ANN$_{15}$**) were then used to predict the outcome in the test set of 1000 patients and the accuracy of these 3 methods in outcome prediction were compared.

Statistical methods

The Hosmer-Lemeshow statistic was used to study the calibration (the accuracy of the system in predicting group outcomes over the entire range of outcome risks) [7]. The area under receiver operator characteristic (ROC) curve was used to assess the ability of the system to distinguish between individual patients who lived and those who died [8,9].

RESULTS

We studied 2962 adult ICU patients. Of these cases, 1000 cases were randomly selected and kept aside as the test cases. Using data from the remaining 1962 cases, the ANN was built taking 22 variables. The entropies (information gain) of all the variables is given in graph and table I.

Figure 1 Graphic representation of variables and their entropies/information gain obtained from 1962 ICU patients.

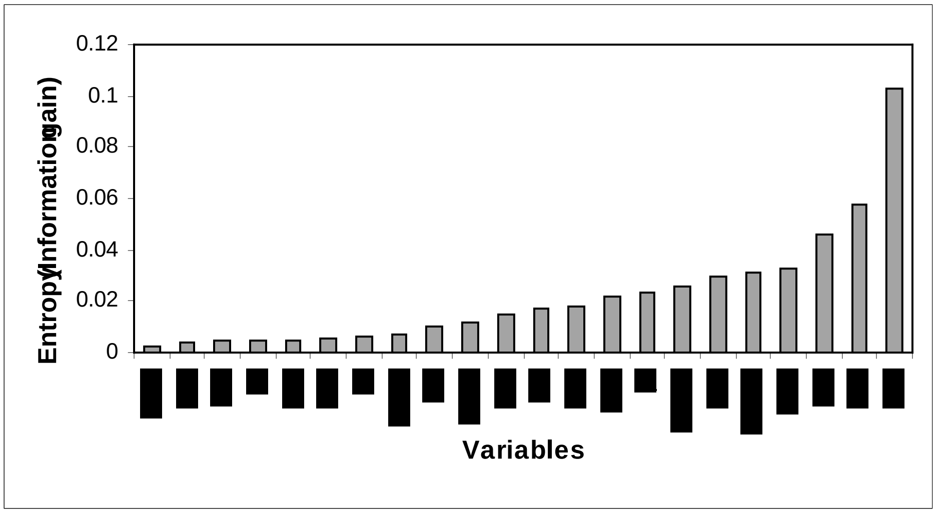


Table 1 Variables arranged in the descending order according to the entropy/information gain.

1	Glasgow Coma Score	12	Acute Renal Failure
2	FIO2	13	Temperature
3	Partial Pressure of Oxygen	14	Hematocrit
4	Mean Arterial Pressure	15	EMOP Emergency surgery
5	Respiratory rate	16	Chronic respiratory disease
6	Age	17	Chronic renal disease
7	Ventricular Rate	18	Potassium
8	Arterial pH	19	Chronic liver disease
9	WBC Count	20	Chronic cardiovascular disease
10	Creatinine	21	Chronic immunological disease
11	Sodium	22	ELOP Elective surgery

From these data it can be inferred that Glasgow Coma Score is the factor that would best predict prognosis, followed by the FIO2, PO2 and so on.

As a second stage we then eliminated the 7 variables with least information gain and included only the top 15 variables according to their information gain. Another ANN model was built with these 15 variables using data from the same 1962 patients.

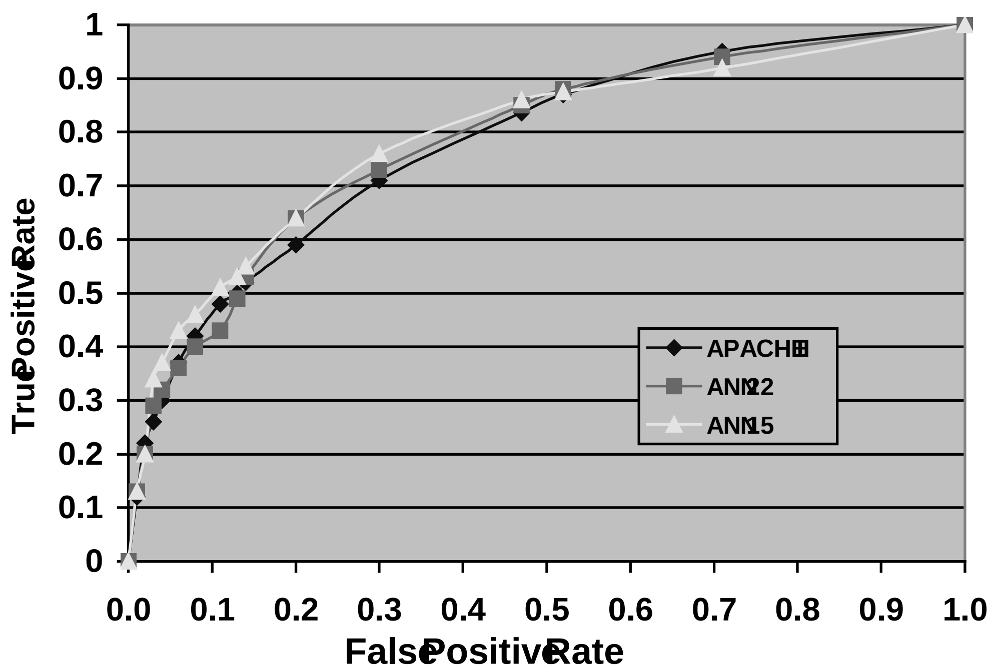
All three models, namely the APACHE II, ANN with 22 input variables (**ANN$_{22}$**) and ANN with 15 input variables (**ANN$_{15}$**) were then used to predict the outcome in the test set of 1000 patients and the accuracy of these 3 methods in outcome prediction were compared.

Table II : Data of 1000 patients (test set) grouped in deciles of risk of death, along with the predicted and actual deaths in each category using all 3 prediction models.

Risk of death	APACHE II Total	Deaths Obs	Deaths Pred	ANN22 (all 22 variables) Total	Deaths Obs	Deaths Pred	ANN15 (15 variables) Total	Obs	Pred
0-10	361	44	12.94	215	20	12.71	221	26	12.60
10-20	210	54	28.63	190	32	25.91	179	22	24.28
20-30	104	41	24.90	141	40	35.34	151	34	36.52
30-40	90	35	30.92	100	28	35.19	101	40	34.99
40-50	73	39	32.52	86	40	38.64	88	38	39.38
50-60	38	24	21.14	57	31	30.76	52	22	28.08
60-70	41	28	26.84	93	48	59.92	73	40	46.90
70-80	36	31	27.19	50	33	37.61	55	48	40.77
80-90	36	31	30.41	41	39	34.40	44	34	37.27
90-100	11	10	10.41	27	26	25.31	36	33	33.84
Total	1000	337	245.9	1000	337	336	1000	337	335
Lemeshow – Hosmer statistic	χ^2=23.5 df=8			χ^2=22.4 df=8			χ^2=27.7		

From table II, the χ^2 value for the ANN22 was 22.4 and ANN15 was 27.7, significantly less than the APACHE II value of 23.5. There was no significant difference between the values of ANN_{22} and ANN_{15}.

Figure 2 Receiver Operating Characteristic (ROC) curves of mortality prediction using APACHE II, ANN22 and ANN15 for 1000 ICU patients



The area under the ROC curve for APACHE II was 77%. This was significantly less compared to ANN_{22} which was 87% (p<0.002) and for ANN_{15} it was 88% (p<0.001), suggesting that the ANN models were able to distinguish between survivors and nonsurvivors more reliably than APACHE II..

DISCUSSION

We found that the performance of the two models of Neural Networks was significantly superior to that of APACHE II model when applied to the given dataset. The Lemeshow-Hosmer statistic showed a better overall goodness-of-fit with the ANN models as compared to the APACHE II model. This statistical test is used to compare prediction of group outcomes over the entire range of predicted outcomes [1,7]. The Receiver Operating Characteristic (ROC) Curve areas for the ANN models showed better discriminative capabilities as compared to APACHE II model. In other words, the ANN models were superior to APACHE II in predicting whether an individual patient would survive or die [1].

There are several possible explanations for the apparent superiority of the ANN over the APACHE II model. Firstly, the APACHE II mortality prediction equation is derived from multiple regression technique which assumes linearity, i.e. an increasing deviation from the midpoint of the normal range corresponds to an increased risk of mortality. This is essentially not the case clinical medicine. In the prediction of mortality using multiple variables, relationship between the different variables (physiological parameters) and the risk of mortality may be complexly non-linear. The neural networks are very good building such non-linear models as they are essentially mapping techniques assigning different risks to different areas of a multidimensional space. While some studies have shown have shown neural networks superior to linear regression models for some medical problems [10-12], others have shown no significant differences between the linear regression and neural network models [13]. One such study by Wong and Young compared prediction by APACHE II with that by ANN in patients admitted to ICU in UK. This study showed that there was no significant difference between the 2 approaches for predicting survival [13].

Secondly, the APACHE II model has been primarily derived based on western cohort that is not representative of the ICU patient population of the Indian sub-continent [5]. There is significant difference in the case-mix in Indian and American ICUs. Indian ICU patients also differ from American and European ICU patients with respect to other factors which may influence outcome, including lead-time bias and differences in organization and utilization of health care resources [5,14]. Hence the APACHE II score, with its existing weights, does not predict survival reliably in Indian patients [5]. Thus our ANN models may have outperformed the APACHE II system mainly because they were trained using Indian patients. This may also explain why Wong and Young did not find ANNs superior to APACHE II in patients from the UK [13]. The APACHE II system could predict outcome so well in their cases that there was little additional benefit to be obtained from ANNs.

In addition to the difference in the predictive performance of the APACHE II and neural networks, we also found that some of the variables that have been used in the development of model are redundant, do not contribute in improving the accuracy of prediction and could as well be eliminated from the model building process. These variables with the least entropy are mostly the ones that indicate chronic health evaluation/major system status and the variables that are most predictive are the ones that indicate the acute physiological status and vital parameters. The disease category for which patient was admitted to the ICU too was important to predict outcome. Wong and Young too have found that their ANN model did not require these data [13].

While neural networks have several advantages [2,11,13] over traditional statistical techniques, they have several drawbacks too. A major drawback [15] of neural networks is that they are primarily "black-box" models and have limited ability to explicitly identify possible causal relationships. They are more difficult to use in the field and need greater computational resources. Finally the neural network model development process is empirical and there are many methodological issues that remain to be solved.

Despite the drawbacks of neural networks, they remain promising and robust tool for developing models for risk stratification. The accuracy of the neural network model can be further improved by increasing the size of the training data. There is also a need for developing a linear regression model on the lines of APACHE for the Indian ICU population which is representative of the tropics since the case mix and other factors that might have significant impact on the mortality of this group of patients have not been accurately represented and modeled in any of the APACHE models. Only after it is developed that we could accurately compare the accuracy of prediction of mortality by the linear regression and neural network models built on samples of similar patient population.

REFERENCES

1 Lemeshow S, Le Gall J. Modeling the severity of illness in ICU patients – a systems update. JAMA 1994; 272: 1049-1055.

2 Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. Lancet 1995; 364: 1075-1079.

3 Hinton GE. How neural networks learn from experience. Sci Am 1992; 267: 145-151.

4 Angus DC, Sirio CA, Clermont G, Bion J. International compar5isons of critical care outcome and resource utilization. Crit Care Clin 1997; 13: 389-407.

5 Parikh CR, Karnad DR. Quality, cost and outcome of intensive care in a public hospital in Bombay, India. Crit Care Med 1999; 27: 1754-1759.

6 Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II – a severity of disease classification system. Crit care med 1985; 13: 818-829.

7 Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982; 115: 92-106.

8 Hanley JA, Mcneil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29-36.

9 Hanley JA, Mcneil BJ. A method of comparing the areas under the receiver operating characteristic curves derived from the same cases. Radiology 1983; 148: 839-843.

10 Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 1991; 115: 843-848.

11 Katz S, Katz AS, Lowe N, Quijano RC. Neural net – bootstrap hybrid methods for prediction of complications in patients implanted with artificial heart valves. Heart Valv Dis 1994; 3: 49-52.

12 Patil S, Henry JW, Rubenfire M, Stein PD. Neural network in the clinical diagnosis of acute pulmonary embolism. Chest 1993; 104: 1685-1689.

13 Wong LS, Young JD. A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks. Anaesthesia 1999; 54: 1048-1054.

14 Sachdeva RS, Guntupalli KK. International comparisons of outcomes in intensive care units. Crit Care Med 1999; 27: 2032-2033.

15 Blackstone EH. Current state of risk factor analysis. Heart Valv Dis 1994; 3: 45-48.