Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression

Gordon S. Doig[1], Kevin J. Inman[2], William J. Sibbald[2], Claudio M. Martin[2], and James McD. Robertson[1]

[1]Department of Epidemiology and Biostatistics, University of Western Ontario,
[2]The Richard Ivey Critical Care Trauma Centre, Victoria Hospital Corp.,
London, Ontario, Canada

ABSTRACT

The objective of this study was to compare and contrast two techniques of modeling mortality in a 30 bed multi-disciplinary ICU; neural networks and logistic regression. Fifteen physiological variables were recorded on day 3 for 422 consecutive patients whose duration of stay was over 72 hours. Two separate models were built using each technique. First, logistic and neural network models were constructed on the complete 422 patient dataset and discrimination was compared. Second, the database was randomly divided into a 284 patient developmental dataset and a 138 patient validation dataset. The developmental dataset was used to construct logistic and neural net models and the predictive power of these models was verified on the validation dataset. On the complete dataset, the neural network clearly outperformed the logistic model (sensitivity and specificity of 1 and .997 vs. .525 and .966, area under ROC curve .9993 vs. .9259), while both performed equally well on the validation dataset (area under ROC of .82). The excellent performance of the neural net on the complete dataset reveals that the problem is classifiable. Since our dataset only contained 40 mortality events, it is highly likely that the validation dataset was not representative of the developmental dataset, which led to a decreased predictive performance by both the neural net and the logistic regression models. Theoretically, given an extensive dataset, the neural network should be able to perform mortality prediction with a sensitivity and a specificity approaching 95%. Clinically, this would be an extremely important achievement. In future trials, we intend to investigate the performance of an application-specific, state of the art neural network on a more representative, comprehensive prospective patient database.

ROC = receiver operating characteristic

INTRODUCTION

Today, it is common practice to assign a severity-of-illness score to a patient upon entry into the intensive care unit (ICU). Common ICU scoring systems include the Acute Physiology and Chronic Health Evaluation (APACHE), the Mortality Probability Models (MPMs), the Simplified Acute Physiology Score (SAPS) and the Pediatric Risk of Mortality (PRISM) scoring system [1].

The scoring strategies for each of these systems differ markedly, but all combine measures of current physiological status with various preexisting risk factors to produce a surrogate measure of risk or disease severity. This surrogate measure can then be used as a tool to aid in quality assurance, resource allocation, clinical decision making, the evaluation of new therapies and outcome prediction [1,2,3].

In the APACHE and MPM approach to predicting outcomes, the day 1 risk factors identified by each model are entered into a logistic regression equation which then produces a predicted probability of mortality. These logistic regression models usually perform very well when predicting the expected mortality experience of an ICU but fall short of clinical usefulness when predicting mortality of the individual patient [2,4].

Recent research has shown that events occurring after ICU admission are more useful predictors of outcomes than ICU admission status [4]. It has also been shown that if the APACHE II scores remain high in the face of continued maximal intervention, fatal outcome can be predicted [5]. These studies indicate that a scoring system enacted some time after admission should have a better predictive performance than a scoring system enacted at admission.

Some researchers believe that understanding the patterns that are associated with survival or death may require the use of alternative

0195-4210/92/$5.00 © 1994 AMIA, Inc.

mathematical approaches, such as set theory or fuzzy logic, which may ultimately be more fruitful than further attempts at refining existing systems [5, 6].

Alternative techniques such as nonlinear discrete neural networks, have recently begun to be applied to some classical medical problems. These techniques are derived from the engineering disciplines of pattern recognition and signals processing and are extremely promising because they offer the potential for ever-improving performance through dynamic learning [7].

A neural network trained on 351 patients with a high likelihood of myocardial infarction outperformed emergency room physicians when presented with 331 new cases of patients presenting with anterior chest pain. The physicians diagnosed myocardial infarction with a sensitivity and specificity of 77% and 84% respectively, whereas the neural network performed with a sensitivity and specificity of 97.2% and 96.2% [8].

Neural networks have outperformed clinicians on the diagnosis of hepatic masses [9], pulmonary emboli [10], and breast tumors [11]. Artificial neural networks have also shown their potential usefulness in the ICU by predicting the length of ICU stays after cardiac surgery [12].

The purpose of this study was to compare and contrast the performance of a relatively simple back-propagation, associative-learning neural network with a classical multivariate logistic regression approach to predicting ICU mortality based on day 3 physiology scores.

METHODS

Patient Selection

During a six month period from August 5, 1991 to February 5, 1992, 614 patients were admitted to the 30 bed multi-disciplinary adult critical care unit. The only entry criterion for this study was a duration of stay greater than 72 hours. Four hundred and twenty-two patients met this criterion and were therefore eligible.

Data Collection

Fifteen variables were recorded daily for the duration of stay for each study entrant. These variables were identified from the literature [13] and from clinical experience. They were; presence of acute renal failure, packed cell volume, heart rate, FIO_2, serum sodium, PaO_2, pH, respiratory rate, systolic and diastolic blood pressures, serum potassium, temperature, white blood cell count, serum creatinine and the Glasgow coma score. The definition and recording of all variables was consistent with the methods outlined for data collection for the APACHE II scoring system [13].

The outcome of interest was ICU mortality in patients with a duration of stay greater than 72 hours.

During the period of the study, the variables were abstracted from patient records and stored in a central codebook. At the conclusion of the study, the codebook was entered into a spreadsheet program and then transferred into PC SAS® version 6.04[1].

Database Validation

Primary data integrity was verified in PC SAS® with algorithms written to filter out biological impossibilities and obvious data entry transpositions. Any values in conflict with the screening filters were re-entered directly from the study codebook.

Secondary validation was carried out by match-merging the study database with a readily available XENIX based ICU management information system (MIS). This MIS database allowed validation of date of birth, ICU entry date, ICU discharge date and ICU discharge status. Since the MIS database is utilized for billing purposes, its entries are double-verified and seldom in error. Conflicts with the study database were resolved by accepting the MIS database as correct.

Primary and secondary error detection rates were then compared as a means to increase confidence in data integrity.

Scale Selection

Variables such as heart rate can convey different information about clinical interventions, outcomes and risks dependent upon the degree of elevation or depression above or below the normal range. For this reason, all variables except serum creatinine, presence of acute renal failure and the Glasgow coma score were separated into high or low distributions about the median. This resulted in a total of 27 variables.

---

[1]PC SAS® version 6.04, SAS Institute Inc., SAS Circle, PO Box 8000,Cary, NC, 27512-8000, U.S.A.

In order to present the variables to the neural network, scaling was required. Each high or low distribution variable was non-parametrically transformed to the z-scale. All 27 z-transformed variables were presented to the neural network and the logistic model to maintain consistency.

Logistic Regression

Logistic regression was performed using PROC LOGIST, PC SAS®, version 6.04 [14]. Variables were considered as candidates for inclusion in the model based on a univariate logistic regression p-value $\leq$ 0.25 [15]. A multiple-step backward model selection method was used and variables were removed from the model if significance fell above a p-value of 0.10. After the final model was evaluated, first order interaction terms were assessed.

Neural Network

A commercially available back-propagation, associative-learning neural network was used for this simulation[2]. All 27 variables were presented to a 3 layered network with 27 input nodes, 18 hidden nodes and 1 output node. A logistic activation function was used and the output node generated a probability of mortality ranging between 0 and 1.

Through an error-minimization technique known as back-propagation, the neural network optimizes weights between nodes such that important patterns between variables are recognized.

Comparisons

The performance of the neural network and the logistic model were compared under two different conditions. First the neural network and the logistic regression techniques were exposed to the complete database. Their ability to discriminate between patients who lived or died was then compared.

Second, a developmental and validation subset were randomly selected from the complete dataset. The developmental dataset contained 284 patients and was used to create a new logistic model and a new neural network model. The two models were then rated on their ability to discriminate between patients who lived or died in the validation dataset, to which they had never been exposed.

Discrimination was assessed using the area under the receiver operating characteristic (ROC) curve of each model [16]. Performance was also assessed by comparing the sensitivity and specificity of the approaches at the arbitrary classification threshold of 0.5.

RESULTS

Patient Population

The average age of the study subjects was 61 years and the average duration of ICU stay was 7.3 days. For the period of the study, the average day 1 APACHE II score was 25.3. The study population experienced a 9.5% mortality rate.

Database Validation

Primary validation revealed a coding error rate of 2.1 % and independent secondary validation against the MIS database revealed an error rate of 2.4%. Comparison of study database errors against codebook values did not reveal any obvious transposition errors in the codebook. Only 0.2% of values were missing due to initial failure to perform laboratory tests at bedside.

Logistic Regression

Complete Dataset

The final model contained seven significant physiological variables. They were; presence of acute renal failure (ARF), high serum sodium, high pH, high diastolic blood pressure (DBP), the Glasgow coma score (GCS), high $PaCO_2$, and low serum sodium (see table 1).

Table 1	β	SE	OR	p
Intercept	-1.6136	0.98	0.199	0.101
ARF	2.3518	0.59	10.50	0.000
Hi Na	0.9741	0.28	2.648	0.000
Hi pH	0.9659	0.39	2.627	0.014
Hi DBP	-1.1326	0.38	0.322	0.003
GCS	-0.2390	0.06	0.787	0.000
Hi $PaCO_2$	-0.8835	0.49	0.413	0.076
Low Na	1.6940	0.40	5.441	0.000

β=regression parameter estimate, SE=standard error of regression parameter, OR=odds ratio, p= p-value

At a classification threshold of 0.5, the logistic regression model performed with a sensitivity of .525 and a specificity of .966. The positive predictive value was .618 and the negative

[2]NeuroShell™,Ward Systems group, Inc., 245 W. Patrick St., Frederick, MD 21701, U.S.A.

predictive value was .951. The area under the ROC curve was .9259.

Validation Dataset

The developmental dataset of 284 patients produced a final model containing seven different significant variables. The seven variables were; high serum sodium, high diastolic blood pressure, GCS, high PaCO2, low serum sodium, low serum potassium, and low temperature (see table 2).

Table 2	β	SE	OR	p
Hi Na	1.2876	0.34	3.624	0.000
Hi DBP	-1.8092	0.47	0.164	0.000
GCS	-0.2357	0.06	0.790	0.000
Hi PaCO2	-1.0222	0.50	0.360	0.042
Low Na	1.8762	0.64	6.529	0.003
Low K	2.0746	0.80	7.961	0.009
Low Temp	0.4771	0.22	1.611	0.029

At a classification threshold of 0.5, this model performed with a sensitivity of .133 and a specificity of .976. The positive predictive value was .400 and the negative predictive value was .902. The area under the ROC curve was .8320.

Neural Network

Complete Dataset

The network converged on a solution after 15,837,750 iterations. This took 17:10:43 hours on a 27 MHz 386.

At a classification threshold of 0.5, the neural network performed with a sensitivity of 1.0 and a specificity of .997. The positive predictive value was .976 and the negative predictive value was 1.0. The area under the ROC curve was .9993.

Validation Dataset

Using the 284 patient developmental database, the network converged on the optimum predictive solution after 20,300 iterations, which took 17:21 minutes on a 27 MHz 386.

At a classification threshold of 0.5, the neural network classified the 138 patient validation database with a sensitivity of .267 and a specificity of .976. The positive predictive value was .571 and the negative predictive value was .916. The area under the ROC curve was .8178.

DISCUSSION

Back-propagation neural networks have traditionally excelled at classification (pattern recognition) problems. They are most useful in situations where the relationship between the input and the output is nonlinear and training data are abundant [17].

On the complete dataset, the back-propagation network clearly outperforms logistic regression with respect to the classification of mortality and survivability (sensitivity and specificity of 1.0 and .997 verses .525 and .966).

With the neural network performing with an area under the ROC curve of .9993 and only one misclassified event, we can conclude that the 15 recorded day 3 physiological variables adequately describe the mortality patterns experienced over the period of the study.

Baxt's neural network predicted myocardial infarction by placing diagnostic importance on clinical variables that have not previously been shown to be highly predictive for infarction [18]. Since the etiology of mortality is much more complex than the etiology of infarction, and since discrimination was so successful with our neural network, it suggests that patterns and predictors of mortality are being detected that were not detected using the traditional logistic regression approach.

When 2/3 of the complete dataset was used for model building and 1/3 for validation, the overall predictive performance of the two approaches was identical (area under ROC = .82). The neural network was more sensitive over the range of decision thresholds while the logistic model was more specific.

Performance of any predictive model on a validation dataset depends on how representative the validation cases are of the developmental cases and on how well the model can classify the developmental dataset. Theoretically, if the validation dataset is truly representative of the developmental dataset, then the predictive performance will approach the level of developmental classification.

If the neural network were exposed to more cases, in the form of a larger dataset, there is no reason to suspect that a similar level of classification would not occur. If this dataset were extensive enough to cover most patterns of mortality, then a predictive sensitivity and specificity of over 95% could reasonably be expected.

Clinically this would be an extremely important achievement. Improved predictive performance would enhance quality assurance, resource allocation, and the evaluation of new therapies. With a sufficiently high predictive performance, the neural network would also be an unprecedented ancillary aid in clinical decision making at the individual level.

Primary research in neural networks is a dynamic and rapidly progressive field. We intend to investigate the performance of an application-specific, state of the art neural network on a more representative, comprehensive prospective patient database.

Reference

[1]. Seneff M, Knaus WA. Predicting patient outcome from intensive care: a guide to APACHE, MPM, SAPS, PRISM and other prognostic scoring systems. J Intensive Care Med 1990;5:33-552

[2]. Rutledge R, Fakhry SM, Rutherford EJ, Muakkassa F, Baker CC, Koruda M and Meyer AA. Acute Physiology and Chronic Health Evaluation (APACHE II) score and outcome in the surgical intensive care unit: an analysis of multiple intervention and outcome variables in 1,238 patients. Crit Care Med 1991;18:1048-1053

[3]. Knaus WA, Wagner DP, Draper EA, Simmerman JE, Bergner M, Bastos PG, Sirion CA, Murphy DJ, Lotring T, Damiano A and Harrel FE. The APACHE III Prognostic System; Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991;100:1619-1636.

[4]. Ferraris VA, Propp ME. Outcome in critical care patients: A multivariate study. Crit Care Med 1992;20:967-976.

[5]. Civetta JM, Hudson-Civetta JA, Kirton O, Aragon C and Salas C. Further appraisal of APACHE II limitations and potential. Surg Gynecol Obstet 1992;175:195-203.

[6]. Chang RWS, Jacobs S and Lee B. Predicting outcome among intensive care unit patients using computerized trend analysis of daily Apache II scores corrected for organ system failure. Intensive Care Med 1988;14:558-566.

[7]. Levin E, Gewirtzman R, Inbar GF. Neural network architecture for adaptive system modeling and control. Neural Networks 1991;4:2:185-191.

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

[9]. Maclin PS and Dempsey J. Using an artificial neural network to diagnose hepatic masses. J Med Syst 1992;16: 215-225.

[10]. Scott JA and Palmer EL. Neural network analysis of ventilation-perfusion lung scans. Radiology 1993;186:661-664.

[11]. Goldberg V, Manduca A, Ewert DL, Gisvold JJ, Greenleaf JF. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. Med Phys. 1992;19:1475-1481.

[12]. Tu JV and Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Proc Annu Symp Comput Appl Med Care 1992; 666-672.

[13]. Knaus WA, Draper EA, Wagner DP, and Zimmerman. APACHE II: A severity of disease classification system. Crit Care Med 1985;13:818-830.

[14]. SAS® Institute Inc., SAS® User's Guide: Statistics, Version 6, 4th Edition, Cary, NC: SAS® Institute Inc. 1990.

[15]. Hosmer DW and Lemeshow S. Applied Logistic Regression. John Wiley and Sons, New York, 1989.

[16]. Erdreich LS and Lee ET. Use of relative operating characteristic analysis in epidemiology. Am J Epidemiol 1981;114:649-662.

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

[18]. Baxt WG. Analysis of the clinical variables driving decision in an artificial neural network trained to identify the presence of myocardial infarction. Ann Emerg Med 1992;21 :1439-1444.