Artificial Neural Networks to Predict Mortality in Critical Care Patients: An Application of Supervised Machine Learning.

DAVID A. COOK, PhD, FJFICM, FANZCA
Intensive Care Unit, Princess Alexandra Hospital, Brisbane
School of Information Technology and Electrical Engineering, University of Queensland

David Cook is an Intensive Care specialist at the Princess Alexandra Hospital, Brisbane. His PhD was awarded for his work in Information Technology and Electrical Engineering. His research is directed towards the integration of contemporaneous, formative computer analysis into the delivery and assessment of patient care, with interests in statistical modelling, data integration and operations research.

Outcome prediction in intensive care is a challenging process. It requires accurate synthesis of quality data and application of prior experience to the analysis. To facilitate this process, artificial neural network (ANN) technology is being increasingly used. ANNs are a form of artificial intelligence capable of analysing complex medical data. They are a class of models and learning methods that superficially resemble the interconnecting neuronal architecture of the human brain. "Learning" occurs with iterative changes in the interrelationships between the "neurones".

There is a considerable amount of activity in the application of ANNs to the medical area. A Medline® search of "artificial neural networks" currently picks up a list of over a thousand papers. In a recent review[1] of artificial intelligence in the Intensive Care Unit (ICU), only six paragraphs described ANN applications in the ICU. This review will provide more background and detail, and will focus on applications of ANNs in ICU to predict patient death and resource use. There are two aims. The first is to describe the machine learning method of ANNs. The second is to review applications of ANNs in the ICU, with particular reference to the estimation of probability of death of adult ICU patients and the length of ICU stay.

Machine Learning Applied to Classification and Regression Problems

Machine learning algorithms use iterative adjustment to parameters to train a model. An error function that compares the outcomes of a training dataset to model predictions is minimised to optimise the model performance. ANNs have been applied to modelling the outcomes of ICU patients. This is an example of a supervised learning task where training datasets of patient related and diagnostic variables, when corresponding resultant patient outcomes are known. Though these approaches are

computationally intensive, it is practical to model ICU patient outcomes in this way because of the ready availability of powerful desktop processors.

1. Conceptual Framework of ANNs

ANNs provide a diverse range of models for classification and regression problems. They have architectures of networks of interconnected simple processors. The training process involves learning patterns in the training data, often by modifying the weights that link the interconnected processing units. The following discussion will be limited to the supervised learning instance. A more detailed introduction to this topic is available in other references.[2-5] The most commonly used example of an ANN is the multilayer perceptron (MLP) which is shown in Figure 1.

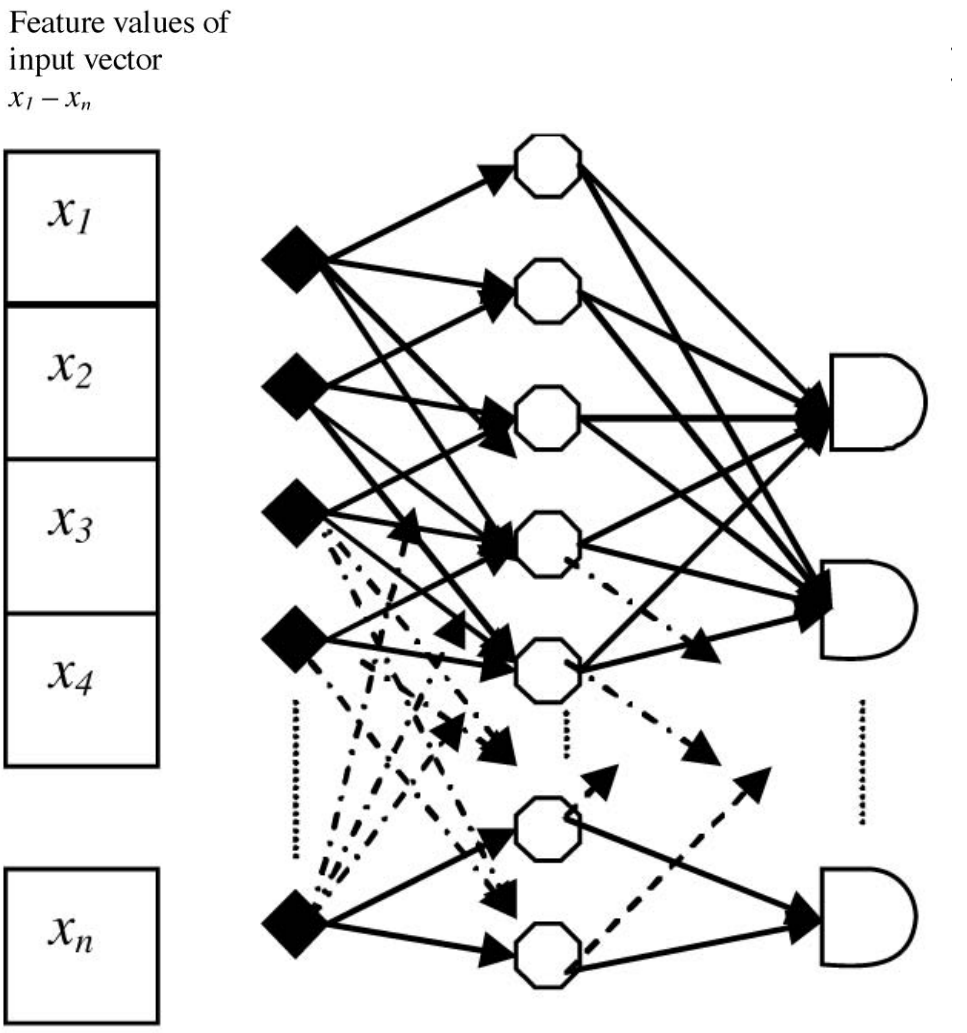

Figure 1. Architecture of 2 Layer Multi-Layer Perceptron. ◆ Input units, ○ Hidden units in hidden layers, → Weights, ⊐ Output units.

This example of a 2 layer MLP has an input layer of n input units (left, ◆) equal to the dimension of the input vector. It has hidden layer units (centre, ○) and output layer units (right, D). The data input is an n dimensional vector, $\mathbf{x}$. In the example, feature values x_1 to x_n are presented at the input nodes. The arrows represent connecting weights. Each input node connects to each node in the layer below. The vector of weights at each node are $\mathbf{w}$. In the hidden layer in Figure 1, $\mathbf{w}$ will have n elements. The input to each hidden unit is the sum of weighted inputs. This is the sum of elements of the dot product of the vectors $\mathbf{w}$ and $\mathbf{x}$ i.e. $w_1x_1+w_2x_2+\ldots\ w_nx_n$ and is abbreviated as $\sum \mathbf{w}\cdot\mathbf{x}$. The input is processed by a bounded increasing activation function to produce a non-linear output. A commonly used activation function is the sigmoid function $f(\sum \mathbf{w}\cdot\mathbf{x}) = \frac{1}{1+e^{\sum w\cdot x}}$.

The outputs of the processing units are connected by weights to the next layer. The next layer in this example is the output layer, though further hidden layers are possible. At the output unit, the sum of the weighted signals is processed into the output signal.

Supervised learning proceeds by iterative adjustments to the inter-connecting weights of the ANN, to minimise the sum of the squared errors between observed and predicted output values. Initially, weights are set to random values. Subsequent adjustment of the interconnecting weights in the network is by gradient descent down the error surface seeking a minimum error value. Back propagation is one such algorithm whereby the weights in the network are adjusted, according to the derivative of the error with respect to the weights.[5]

MLP ANNs have a number of potential issues[5] which are common to many numerical optimisation procedures. For example, the gradient descent algorithm can become trapped in local error minima, rather than proceeding to the optimal solution. The MLP can produce different training results depending on the starting weights, by terminating at different local error minima. As well, MLPs are prone to over-fitting. It is usual practice to monitor performance on an external verification data set during the training. The training process is terminated when the generalisation performance begins to deteriorate. In addition, a test dataset is used to assess the model performance. Whilst there are strategies to limit the problems caused by these issues, it is important to train many MLPs and to choose the best ANN for the application at hand.

A suitable approach is to divide the available data into three sets for each trial. A training set is used to train the ANN. A verification set is used to monitor the generalisation performance so that training can be stopped before over-fitting of the model. Usually, an ANN will improve predictions with exhaustive training until nearly all examples in the training data are correctly predicted. However, this often results in a loss of ability to generalise to unseen data, so the verification set is used to stop adjustments when overtraining occurs. A test set is used to assess the model performance on the data not used for model development.

To illustrate an example of ANN use in the ICU, the patient outcome of interest (survival/death) or a continuous outcome measurement (say, length of stay), and explanatory variables that capture factors which influence patient outcome can be used. Typically measurements that capture patient related factors (physiological measurements, laboratory tests, co-morbidities and age), disease and diagnostic factors (such as acuity or diagnostic group or procedure), and process factors (such as lead

time, referral source or readmissions status) are used as the input vectors. The training process will incorporate and weight the importance of the variables, based on the experience accumulated in the training set.

2. Review of the Use of ANN in the Intensive Care Unit

ANNs are widely applied in the medical area. Lisboa[6] provides a thorough overview of medical applications. Non-ICU applications include diagnosis,[7-13] prognosis,[14-20] physiological and laboratory data interpretation[9, 12, 21-23] and pharmacology.[12, 24] There are a number of reports describing ANNs used to model ICU patient data and outcomes. The following review of applications summarises the use of ANNs in the ICU from the perspective of predicting mortality or resource use.

Two studies have examined cardiac surgical sub-sets of ICU patients. Lippmann and Shahian[25] used a MLP ANN, logistic regression and Bayesian analysis to model the survival outcomes of 80,606 cases. Fifty-nine demographic, physiological, laboratory, diagnosis and cardiac assessment features were collected, of which 36 were used in the model. The calibration of the logistic regression model was the best, but the model had an area under the ROC curve of only 0.76. Orr[26] reported an ANN to estimate the risk of death in cardiac surgical patients using only seven variables selected from a patient database. This model had good calibration, but lacked discrimination, with the area under the ROC curve of only 0.74.

Buchmann et al[27] compared a logistic regression model, MLP, GRNN (generalised regression NN) and a probabilistic neural network to classify ICU patients on the basis of chronicity (length of stay in ICU >7 days), rather than to predict mortality. He found that the discrimination and calibration of the ANNs were superior to logistic regression. Another study of prediction of resource use, as measured by hospital length of stay, was published by Mobley et al.[28] An ANN was trained to estimate the hospital length of stay of 557 coronary care patients using 74 variables including demographic characteristics, physiological observations, laboratory results, diagnostic tests and index events. The ANN was able to predict length of stay within 24 hrs in 72% of patients. However, this was only marginally better than using the mean length of stay of each diagnostic class in the dataset.

Doig et al[29] compared the predictions of an ANN to a logistic regression model for the classification of a small series of 422 patients. Each patient had already survived to 72 hours in ICU. Variables were selected from the APACHE II system and re-modelled. Remarkable discrimination was achieved on the training set (area under the ROC curve=0.99). The discrimination was less good on the validation set (area under the ROC curve 0.82) suggesting overtraining of the ANN with over-fitting to the training data at the expense of generalisation performance.

Dybowski et al[30] compared an ANN trained with a genetic algorithm to a logistic regression model for predicting the outcomes of a small subset of ICU patients (258 patients) with systemic inflammatory response syndrome. A classification tree and logistic regression were used to select variables from physiological and demographic variables, and index events that occurred during the hospital stay. The ANN had better discrimination than logistic regression (area under the ROC curve 0.86 v 0.75). No assessment was made of the calibration of the models.

Frize et al[31] trained MLPs to classify non-operative (608) and operative (883) ICU patients according to the predicted duration of mechanical ventilation. These models were only assessed on the developmental dataset predictions. By pruning the number

of input features from 51 to 6, classification performance improved and the network complexity was reduced. This study demonstrated a practical approach to limiting network complexity and provides useful documentation of a successful approach to processing and transformation of patient data. However, the ANNs in this study were designed to estimate resource use rather than to predict mortality outcome. A weakness of the study was that the models were not assessed on a separate test dataset, so no conclusions can be drawn about the reproducibility of the modelling, or of the model's generalisation performance.

Two studies compared ANNs to the APACHE II system. Wong and Young[32] compared MLPs to the APACHE II system on 8796 patient admissions collected for an APACHE II database. Both the MLP ANN and the APACHE II system had similar discrimination (area under the ROC curve 0.82-0.84) and calibration. Nimgaonkar et al[33, 34] also compared the performance of ANNs to the APACHE II system to predict mortality in an Indian ICU. A series of 2962 cases were modelled using the input variables for the APACHE II system. They analysed the contribution of each of the features to the models. Discrimination by the ANN was superior to the APACHE II (area under the ROC curve 0.88 v 0.77). The ANN displayed better calibration than the APACHE II model.

Clermont et al[35] used logistic regression and ANN to model hospital mortality outcome on 1647 ICU patients. The demographic and physiology variables were collected under the rules of the APACHE III system. This is quite a good study and should be considered in some detail. In Clermont's study, the component variables of the APACHE III model and the APACHE III score were used to model the probability of patient death. The areas under the ROC curves were in the range of 0.8 (logistic regression) — 0.836 (ANN with coded APACHE III observations). All the models had reasonable calibration when 800 or more cases were used to develop the model. The ANN and the logistic regression models were able to successfully predict ICU patient death. A further important conclusion was that both the logistic regression and ANN model performance deteriorated when the model development set size was reduced below 800 cases. In a real application, at a busy hospital ICU like that at the Princess Alexandra Hospital, this requires 5 months of patient data collection for model building. For smaller ICUs, it may represent 2-3 years of data collection.

Equally important was the author's practical choice of the size of 447 cases for the test dataset for model assessment. The size of this assessment set is a trade-off between the expediency of collecting patient data and the important statistical issues of the power and precision of the model assessment. For the PAH ICU and other busy ICUs, 447 patients requires about 3-4 months of data collection. For smaller units, a year may be required just to collect enough patient data for model assessment. Balanced against this, is the issue of smaller assessment datasets giving low statistical power to the Hosmer-Lemeshow *C* (H-L *C*) test to detect imperfect calibration, and the loss of statistical precision in estimating the area under the ROC curve. Clermont et al made their choice based on a timed period of data collection. The size of the dataset they used strikes a reasonable compromise between the time for data collection and statistical issues.

There are limitations to Clermont's study. As Paetz[36] in a letter to the Editor in a subsequent edition of *Critical Care Medicine* commented, the study did not involve resampling to demonstrate the robustness or reproducibility of the approach. It is possible that the non-random split of cases to the training and validation sets was a

major determinant of the model's reported performance. Replicates of the modelling on alternative random data selections and re-sampling to provide alternative assessment sets are necessary to demonstrate consistency of the modelling approach. I would add to these criticisms, that the authors used consecutive patients to build the models. The last 447 consecutive cases in the dataset were used to assess all the models. Even when smaller training sets were explored, non-random, consecutive sampling was used. Introduction of bias into the model, the effects of influential outliers, or fortuitous sampling cannot be excluded with their methodology.

There are further technical limitations imposed by their methods. The patients' diagnoses or diagnostic coding were not included in the variables for the model. Also, the authors relied on the APACHE III algorithm for the weights that they used to pre-process the variables for both the MLP ANN and the logistic regression model. These APACHE III weights are added together to give the APACHE III score, which the authors also selected as a variable in some models. A lack of diagnosis variables and reliance on the APACHE III system may have limited the quality of the models that could be built. The authors did not record how many ANNs were trained to yield the optimal performance ANN, so it is not clear whether a limited or an exhaustive survey of possible ANNs was conducted. Despite these short-comings, this paper is probably the best example in the critical care literature.

Conclusion

Several applications of ANNs to prediction of ICU mortality have been reviewed. Overall, the performances of ANNs on ICU mortality prediction tasks appear as good as or better than logistic regression, on the datasets on which the models have been developed. Conclusions about the generalisation of these ANN models outside the contexts where each was developed can only be made when such models are applied and validated more widely. Variation in data quality and measurement may be the limitation to model performance, rather than the nature of the statistical tool.

Bibliography

1. Hanson WC, Marshall BE. Artificial intelligence applications in the ICU. *Crit Care Med* 2001; 29:427-435.
2. Maren A, Harston C, Pap R, eds. Handbook of Neural Computing. San Diego: Harcourt Brace Jovanovich, 1990.
3. Statistica Neural Networks. Tulsa, Oklahoma: Statsoft, 1998.
4. Bishop C, ed. Neural networks and Machine Learning. Berlin: Springer Verlag, 1998.
5. Reed RD, Marks RJ. Neural Smithing: Supervised learning in feedforward artificial neural network. 1st ed. Cambridge, Massachusetts: MIT Press, 1999.
6. Lisboa PJG. A review of evidence of health benefit from ANNs in medical intervention. *Neural Networks* 2002; 15:11-39.
7. Floyd CE, Lo JY, Yun AJ et al. Prediction of breast cancer malignancy using an artificial neural network. *Cancer* 1994; 74:2944-2948.
8. Ortiz J, Ghefter CGM, Silva CES et al. One year mortality prognosis in heart failure: A neural network approach based on echocardiographic data. *JACC* 1995; 26:1586-1593.
9. Selker HP, Griffin JL, Patil S et al. A comparison of performance of mathematical predictive methods for medical diagnosis: Identifying acute cardiac ischaemia among emergency department patients. *J Investigative Med* 1995; 43:468-476.
10. Doyle HR, Parmanto B, Munro WP et al. Building clinical classifiers using incomplete observations — A neural network ensemble for hepatoma detection in patients with cirrhosis. *Meth Inform Med* 1995; 34:253-258.
11. Setiono R. Extracting rules from pruned ANN for breast cancer diagnosis. *AI in Med* 1996; 8:37-51.

12. Itchhaporia D, Snow PB, Almassy RJ et al. ANN: Current status in cardiovascular medicine. *JACC* 1996; 28:515-521.
13. Eisenstein EL, Alemi F. A comparison of 3 techniques for rapid model development: An application in patient risk stratification. *Proc Med Informat Ass* 1996:443-447.
14. Lette J, Colletti BW, Cerino M et al. Artificial intelligence vs logistic regression statistical modelling to predict cardiac complications after non — cardiac surgery. *Clin Cardiol* 1994; 17:609-614.
15. Doyle HR, Dvorchik I, Mitchell S et al. Predicting outcome after liver transplantation. *Ann Surg* 1994; 219:408-415.
16. Hamamoto I, Okada S, Hashimotot T et al. Prediction of the early prognosis of the hepatectomised patient with hepatocellular carcinoma with a neural network. *Comput Biol Med* 1995; 25:49-59.
17. Dombi GW, Nandi P, Saxe JM et al. Prediction of rib fracture outcome by an artifical neural network. *J Trauma, Infection and Critical Care* 1995; 39:915-921.
18. Dvorchik I, Subotin M, Marsh W et al. Performance of multi-layer feedforeward neural network to predict liver transplantation outcome. *Meth Inform Med* 1996; 35:12-18.
19. Izenberg SD, Williams MD, Luterman A. Prediction of trauma mortality using a neural network. *American Surgeon* 1997; 63:275-281.
20. Jefferson MF, Pendleton N, Lucas SB et al. Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with non-small cell lung carcinoma. *Cancer* 1997; 79:1338-1342.
21. Reibnegger G, Weiss G, Werner-Felmayer G et al. Neural network as a tool for utilizing laboratory information: Comparison with linear discriminant analysis and with classification and regression trees. *Proc Natl Acad Sci USA* 1991; 88.
22. Forsstrom JJ, Dalton KJ. Artificial neural network for decision support in clinical medicine. *Ann Med* 1995; 27:509-517.
23. Jorgensen JS, Pedersen JB, Pedersen SM. Use of neural network to diagnose acute myocardial infarction. *Methodology Clin Chem* 1996; 42:604-612.
24. Brier ME, Aronoff GR. Application of artificial neural network to clinical pharmacology. *Int J Clin Pharm Therapeutics* 1996; 34:510-514.
25. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. *Ann Thor Surg* 1997; 63:1635-1643.
26. Orr RK. Use of a probablistic neural network to estimate risk of mortality after cardiac surgery. *Med Dec Making* 1997; 17:178-185.
27. Buchman TG, Kubos KL, Seidler AJ et al. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical ICU. *Crit Care Med* 1994; 22:750-762.
28. Mobley BA, Leasure R, Davidson L. Artificial neural network predicitons of lengths of stay on a post coronary care unit. *Heart and Lung* 1995; 24:251-256.
29. Doig GS, Inman KJ, Sibbald WJ et al. Modelling mortality in the ICU: comparing the performance of a back propagation, associative learning neural network with multivariate logistic regression. *Proc Ann Sym. Computer Application in Med Care* 1994; 17:361-365.
30. Dybowski R, Weller P, Chang R et al. Prediction of outcome in critically ill patients using artificial neural networks synthesized by genetic algorithm. *Lancet* 1996; 347:1146-1150.
31. Frize M, Ennett CM, Stevenson M et al. Clinical decision support systems for ICU: Using artificial neural networks. *Med Eng Physics* 2001; 23:217-225.
32. 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.
33. Nimgaonkar A, Sudarshan S, Karnad DR. Prediction of mortality in an Indian ICU: Comparison between APACHE II and artificial neural network. (Hansraj Prize paper). Proceedings of the Annual Scientific Meeting, Indian Society of Critical Care Medicine. 2001:43-46.
34. Nimgaonkar A, Karnad DR, Sudarshan S et al. Prediction of Mortality in an Indian ICU: Comparison between APACHE II and artificial neural network. *Int Care Med* 2004; 30:248-253.
35. Clermont G, Angus DC, DiRusso SM et al. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit Care Med* 2001; 29:291-296.
36. Paetz J. Some remarks on choosing a method for outcome prediction (letter). *Crit Care Med* 2002; 30:724