

Modeling Outcome of Mechanical Intervention after Cardiac Surgery

Jill P. Card and Param K. Singh
ABIOMED, Inc.
Cherry Hill Drive, Danvers MA 01923

ABSTRACT: Cardiac surgery patients who require Intra-aortic balloon pump (IABP) intervention after failure to separate from cardiopulmonary bypass have high mortality rates (35-50%). Databases of pre-IABP patient records and outcomes were used to develop neural network models of predicted mortality. Three data sets were developed, differing by the number of variable and records. There were no missing values in any data set. Parameters had to have Chi-square p values < .1. The first data set consisted of four parameters, training n = 282, and test n = 306. The second data set had 12 input parameters and training and test n = 96. The third data set had 21 input parameters, training n = 27 and test n = 26. Feed forward Backpropagation and Fuzzy ARTMAP neural networks were developed, and the best network compared with logistic regression models. The area under the receiver-operating characteristic curves (ROC), and predictive accuracy were used to evaluate models. For the first data set, logistic regression and Backpropagation gave essentially identical results (ROC area = 0.67 ± 0.03 and accuracy $\approx 64\%$). A Fuzzy ARTMAP model based on the second data set did slightly better than logistic regression (ROC area = 0.72 ± 0.05 versus 0.70 ± 0.05 , accuracy 75.6% versus 62.5%). A Fuzzy ARTMAP developed with the third data set (21 parameters) was superior to logistic regression (ROC area = 0.84 ± 0.08 versus 0.57 ± 0.11 , accuracy 85.4% versus 46.9%). Neural networks were better than logistic regression in assimilating larger numbers of variables and their predictive power indicates clinical utility.

INTRODUCTION

Open Heart surgery is a highly successful medical intervention with more than 330,000 annual procedures in the United States. There is, however, a finite mortality associated with this surgery of between 0.5-10%, the exact number being dependent on the complexity of the procedure, patient condition, demographics, etc. In those patients where separation from cardiopulmonary bypass cannot be achieved even the maximal pharmacological intervention, a mechanical heart support device, the Intra-aortic balloon pump (IABP), is inserted into the patient's arterial system. However, approximately 35-50% [Baldwin, Naunheim] of these patients do not survive to hospital discharge. Until recently the IABP represented the limits of mechanical intervention, but recently ventricular assisted devices (VAD) have been approved for use in such situations [Guyton]. Such devices have greater hemodynamic effectiveness than the IABP but are more invasive and complicated. Thus a need exists for predicting the course IABP intervention so that patients likely to have poor outcomes may be considered for VAD support.

Previous methods used for developing such predictive models have been statistical in nature. [Baldwin]. This paper is concerned with the application of neural network techniques to this problem. The data set that has been used for this evaluation was provided by the Texas Heart Institute (THI) and includes data from the Methodist Hospital in Houston. This data had been used previously to develop a logistic regression predictive model [Baldwin].

MATERIAL AND METHODS

All analyses were performed using a medical record database supplied by THI consisting of 322 patient records with a maximum set of 240 input variables and a single outcome variable of survival. Completeness of patient records varied. Consequently, for analyses requiring a larger set of input variables, fewer complete cases were available for model training and testing (validation). Many of these input variables are correlated and not statistically independent. For example, sex and body surface area (BSA) are highly correlated. In addition, a second database of 320 patients and four input variables, from the Methodist Hospital, Houston, was used for prospective validation of one of the analyses.

Three data sets were derived from the database, distinguished only on the basis of predictive variable selection and consequent sample size difference. The predictor variables of the THI patient record data were analyzed for correlation with in-hospital mortality using Chi-square test, Likelihood Ratio Chi-square test, Mantel Haenszel Chi-square test and Fisher's exact test where appropriate, and only those that had p values <0.1 on at least one of these univariate Chi-square tests were considered for modeling. Each data set consisted of input variables and the observed outcome (mortality).

The first data set (SET 1) used the same predictor variables as in the original analysis [Baldwin]. Four variable (age, BUN, need for pacing and sex) were used, resulting in a training set consisting of 282 THI cases. The test set consisted of 306 patient records from the Methodist Hospital.

The second data set (SET 2) consisted of 12 predictor variables (the four in the above set plus BSA, severity of congestive heart failure, New York Heart Association classification, degree of coronary disease, creatinine, CPB duration, weaning with digoxin, time to wean to IABP). Each predictor variable was present in at least 200 records. The final test and training data sets for this model consisted of 96 complete records each.

The third data set (SET 3) consisted of 21 predictor variables (those listed in the above set plus surgical priority, LVEDP, AST, bilirubin, ischemic time, defibrillation attempts, MPAP, renal disease, hepatic disease) with Chi-square p values $<.1$ for at least one of the listed tests but with no minimum available record count requirement. The resulting train and test sets consisted of 27 and 26 complete patient records, respectively.

Models and Analysis

All three sets were analyzed using either a feed forward neural network with backpropagation of error, or Fuzzy ARTMAP neural network, and compared to a main effects Logistics model. The choice of neural network was based on which of the two neural network architectures yielded the best prediction accuracy for each of the data sets.

The backpropagation neural network employed had an input layer consisting of the input variable data, a hidden layer, and output layer consisting of the predicted surgical outcome variable. Each layer was fully connected to the succeeding layer through a series of weight vectors. Several transfer functions were used during network training including the sigmoid, hyperbolic tangent, and sine functions. The function yielding most accurate train/test statistics was used in any reported final model. The update of weights during learning used the Extended

Delta Bar Delta learning rule [Minai] which incorporates learning rate and momentum terms into weight update algorithms to speed convergence and improve minimization of global error.

The Fuzzy ARTMAP neural network architecture was developed by Carpenter, et al [Carpenter 92] for real-time learning and classification of analog maps using internal self-organization. It incorporates fuzzy logic operations of input and supervision vectors into two ART modules. The ARTa and ARTb modules [Carpenter 88] compute internally organized prototype patterns for classifying input and outcome vectors, respectively. The two ART modules are linked by an inter-ART module, F_{ab} which forms predictive associations between the ARTa incoming vector categories and the ARTb outcome categories. When a predictive mismatch occurs at ARTb, F_{ab} match tracking triggers an increase in the vigilance within ARTa, until sufficient discrimination exists among incoming patterns to predict the proper ARTb output response. In this way, the match tracking system of ARTMAP uses only the generalization necessary to predict accurately.

The logistic regression model predicts the probability of an event of a binary outcome variable as a log linear function of the input variables.

The predictive accuracy of models was computed by calculating the sensitivity and specificity based on a cut point of 0.5 for the normalized output.

The ability of a model to distinguish between outcomes was computed using the area under the receiver operating characteristic (ROC) curve [Hanley]. For a viable model this area should exceed 0.70 [Lemeshow]. The ROC curve for logistic regression is well defined [Baldwin]. The ROC computation for backpropagation was based on varying the decision threshold for the normalized output between 0 and 1 and computing the associated sensitivity and specificity. In the case of Fuzzy ARTMAP models, the specificity and sensitivity determination is based on the magnitude of $T_1 - T_2$:

for $T_1 = \max\{T_{ja} \text{ s.t. } T_{kb} = 1\}$, and

$T_2 = \max\{T_{ja} \text{ s.t. } T_{kb} = 2\}$

where T_{ja} = value of the jth node within the self-organizing layer of ARTa module for a given input vector, x , $0 < T_{ja} < 1$;

T_{kb} = kth node value within the ARTb outcome module to which T_{ja} maps;

T_{kb} takes on only the values 1 and 2 in this application, corresponding to death or survival.

Since $-1 < T_1 - T_2 < 1$, the decision threshold can be varied between $[-1, 1]$ for sensitivity/specificity computation.

Finally, for all models, the area under the ROC curve and its standard error were estimated using procedures in the literature [Hanley].

RESULTS

The results of logistic and neural network modeling for the three data sets are summarized in Table 1.

TABLE I

Data Set	Logistic Model				Neural Network			
	Training		Testing		Training		Testing	
	Accuracy (%)	ROC Area (\pm std err)	Accuracy (%)	ROC Area (\pm std err)	Accuracy (%)	ROC Area (\pm std err)	Accuracy (%)	ROC Area (\pm std err)
SET 1 Backprop	66.5	.72 \pm .026	63.7	.68 \pm .032	68.8	.74 \pm .029	65.0	.68 \pm .031
SET 2 ARTMAP	68.8	.80 \pm .045	62.5	.70 \pm .055	87.6	.92 \pm .029	75.6	.72 \pm .054
SET 3 ARTMAP	100	1.0	46.9	.57 \pm .111	100	1.0	85.4	.84 \pm .081

SET 1 consists of the same train and test set used by Baldwin et al [Baldwin] and the overall accuracy of the logistic train and model sets are comparable at 66.5% and 63.7%. The neural network results shown for SET 1 are for a Backpropagation model with 4 input variables plus a bias term, four hidden nodes in a single layer and one output layer node. The network used the Extended Delta Bar Delta learning algorithm and hyperbolic tangent transfer function for both the hidden and output layers. The network was programmed to prune any hidden nodes which would result in no less than 97.5% accuracy; this resulted in a final hidden layer node count of three at the end of 400 K training iterations. The train and validation accuracy's of 68.8% and 65% are essentially identical to those for logistic regression and for this limited set of input variables the neural network does not result in any improvement over logistic regression.

SET 2 consists of 96 data records equally divided between the train and validation sets. A Fuzzy ARTMAP network architecture with 12 input nodes plus Bias term, 20 committed hidden layer nodes, and a single outcome output node was the neural network model. The recode parameter was initially set at $\beta = 0.1$ and then reduced to 0.03 after 1000 iterations. The choice parameter, α , was set at 0.1. The network was trained for a total of 4500 iterations. Logistic regression results are similar to those from SET 1; the additional input variables do not improve the power of regression. However, the train and validation set accuracy values of 87.6% and 75.6% for the Fuzzy ARTMAP are a substantial improvement over the results from SET 1. These improvement is less pronounced for the ROC are and may indicated a slight overtrain situation.

SET 3 had the largest number of input variables, 21, and the smallest number of complete data records (53). The neural network model also had Fuzzy ARTMAP architecture using 21 input nodes, 5 hidden nodes and a single outcome node. The total number of training set iterations was 200, with the recode parameter, β , set to 0.7 and the choice parameter, α , set at 0.01. The logistic model training was 100% accurate due to the large number of variables and small number of cases, but the accuracy for the validation set was substantially lower at 46.9%. The neural network training set exhibited 100% accuracy (as in the logistic case) and the validation set surpassed all other models fit with an 85.4% prediction accuracy. Perhaps of more significance is that the area under the ROC curve increased substantially to 0.84. while for logistic regression it decreased to 0.57.

The standard error of the ROC area curves increased, as expected, with decreasing sample size for all models.

Both the predictive accuracy and ROC area increased for neural networks as the number of variables increased. This improvement may reflect the strength of neural networks to learn in

the presence of noise and highly correlated input variables. In contrast, the logistic model appeared to suffer when input variables increased for precisely the same reason. The logistic model assumes independence of input variables, and is not able therefore to extract the component of unique information found among many highly correlated variables.

DISCUSSION

We have used separate training and testing sets for developing and validating these models, measured model discrimination using the area under the ROC curve [Hanley] and used a common classification measure for predictive accuracy on all models. The number of records in the databases available were too small for model calibration using formal goodness-of-fit tests. Models based on neural networks and other newer modalities are becoming increasingly common in the literature. However, for acceptance in clinical medicine, these models must be tested for robustness and validity by the same measures used for more traditional statistical models such as logistic regression. Except for model calibration we have followed these measures [Lemeshow] and used logistic regression as a bench mark.

The results of these analyses suggest that neural networks can extract more information from real, highly correlated and fuzzy data than logistic regression. In predicting the outcome after cardiac surgery, we need to recognize that models are inherently unable to predict the risks that an individual patient may be exposed to after surgery, thus predictive ability can never be perfect. An overall accuracy of 85%, as measured for Set 3, may represent the achievable limit. However, such accuracy would have substantial clinical value and is a significant improvement over current regression models that demonstrate accuracy of less than 65%.

Acknowledgment

We wish to express our appreciation to the Texas Heart Institute and Dr. Arthur S. Keats for kindly providing the databases used in this study.

REFERENCES

Baldwin RT, Slogoff S, Noon GP, Sekela M, Frazier OH, Edelman SK, Vaughn WK. A Model to Predict Survival at Time of Postcardiotomy Intraaortic Ballon Pump Insertion. *Ann Thorac Surg.* 1993; 55: 908-13.

Carpenter GA, Grossberg S. The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network. *Computer.* March 1988: 77-78.

Carpenter GA, Grossberg S, Markuzon N, Reynolds JH and Rosen DB. Fuzzy ARTMAP: An Adaptive resonance architecture for incremental learning of analog maps. *IJCNN.* June 1992; III: 309-14.

Guyton RA, Schonberger JP, Everts PA, Jett GK, Gray LA Jr, Gielchinsky I, Raess DH, Vlahakes GJ, Woolley SR, Gangahar DM, et al. Postcardiotomy shock: clinical evaluation of the BVS 5000 Biventricular Support System. *Ann Thorac Surg.* 1993; 56(2): 346-56.

Hanley JA, McNeil BJ, The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*. 1982; 143: 29-36.

Lemeshow S, Le Gall JR. Modeling the Severity of Illness of ICU Patients. *JAMA*. 1994; 272: 1049-55.

Minai AA, Williams RD. Acceleration of Back-Propagation through Learning Rate and Momentum Adaptation. *IJCNN*. January, 1990; I: 676-79.

Naunheim KS, Swartz MT, Pennington DG, Fiore AC, McBride LR, Peigh P, Barnett MG, Vaca KJ, Kaiser GC, Willman VL. Intraaortic Balloon pumping in patients requiring cardiac operations. *J Cardiovasc Thorac Surg*. 1992; 104: 1654-61.