A case study of using artificial neural networks for classifying cause of death from verbal autopsy

Andrew Boulle,[a] Daniel Chandramohan[a] and Peter Weller[b]

Background Artificial neural networks (ANN) are gaining prominence as a method of classification in a wide range of disciplines. In this study ANN is applied to data from a verbal autopsy study as a means of classifying cause of death.

Methods A simulated ANN was trained on a subset of verbal autopsy data, and the performance was tested on the remaining data. The performance of the ANN models were compared to two other classification methods (physician review and logistic regression) which have been tested on the same verbal autopsy data.

Results Artificial neural network models were as accurate as or better than the other techniques in estimating the cause-specific mortality fraction (CSMF). They estimated the CSMF within 10% of true value in 8 out of 16 causes of death. Their sensitivity and specificity compared favourably with that of data-derived algorithms based on logistic regression models.

Conclusions Cross-validation is crucial in preventing the over-fitting of the ANN models to the training data. Artificial neural network models are a potentially useful technique for classifying causes of death from verbal autopsies. Large training data sets are needed to improve the performance of data-derived algorithms, in particular ANN models.

Keywords Verbal autopsies, classification, neural networks

Accepted 10 January 2001

In many countries routine vital statistics are of poor quality, and often incomplete or unavailable. In countries where vital registration and routine health information systems are weak, the application of verbal autopsy (VA) in demographic surveillance systems or cross-sectional surveys has been suggested for assessing cause-specific burden of mortality. The technique involves taking an interviewer-led account of the symptoms and signs that were present preceding the death of individuals from their caretakers. Traditionally the information obtained from caretakers is analysed by physicians and a cause(s) of death is reached if a majority of physicians on a panel agreed on a cause(s). The accuracy of physician reviews has been tested in several settings using causes of death assigned from hospital records as the 'gold standard'. Although physician reviews of VA gave robust estimates of cause-specific mortality fractions (CSMF) of several causes of death, the sensitivity, specificity and predictive values varied between causes of death and between populations[1,2] and had poor repeatability of results.[3]

Arguments to introduce opinion-based and/or data-derived algorithm methods of assigning cause of death from VA data are based on both the quest for accuracy and consistency, as well as the logistical difficulties in getting together a panel of physicians to review what are often large numbers of records. However, physician review performed better than set diagnostic criteria (opinion-based or data-derived) given in an algorithm to assign a cause of adult death.[4] One promising approach to diagnose disease status has been artificial neural networks (ANN) which apply non-linear statistics to pattern recognition. For example, ANN predicted outcomes in cancer patients better than a logistic regression model.[5] Duh *et al.* speculate that ANN will prove useful in epidemiological problems that require pattern recognition and complex classification techniques.[6] In this report, we compare the performance of ANN and logistic regression models and physician review for reaching causes of adult death from VA.

[a] London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E, UK.

[b] Centre for Measurement and Informatics in Medicine, City University, London, UK.

Correspondence: Daniel Chandramohan, Infectious and Tropical Diseases Department, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. E-mail: D.chandramohan@lshtm.ac.uk

Methods

An overview of neural networks

Although often referred to as black boxes, neural networks can in fact easily be understood by those versed in regression analysis techniques. In essence, they are complex non-linear modelling equations. The inputs, outputs and weights in a neural network are analogous to the input variables, outcome variables and coefficients in a regression analysis. The added complexity is largely the result of a layering of 'nodes' which provides a far more detailed map of the decision space. A single node neural network will produce a comparable output to logistic regression, where a function will combine the weights of the inputs to produce the output (Figure 1).

Combining these nodes into multiple layers adds to the complexity of the model and hence the discriminatory power. In so doing, a number of elements, each receiving all of the inputs and producing an output, have these outputs sent as inputs to a further element(s). The architecture is called a multi-layer perceptron (Figure 2).

The study population and field procedures of the VA data used in this analysis are described elsewhere.[1] In brief, data were collected at three sites (a regional hospital in Ethiopia, and two rural hospitals in Tanzania and Ghana). Adults dying at these hospitals who lived within a 60-km radius of the institution were included in the study. A VA questionnaire was administered by interviewers with at least 12 years of formal education. The reference diagnoses (gold standard) were obtained from a combination of hospital records and death certificates by one of the authors (DC) together with a local physician in each site. A panel of three physicians reviewed the VA data and reached a cause of death if any two agreed on a cause (physician review).

The method used to derive algorithms from the data using logistic regression models has been described elsewhere.[4] Each subject was randomly assigned to the *train* dataset (n = 410) or *test* dataset (n = 386), such that the number of deaths due to each cause (gold standard) was the same in both datasets. If a cause of death had odd numbers, the extra subject was included in the train dataset. Symptoms (includes signs) with odds ratio (OR) $\geqslant$2 or $\leqslant$0.5 in univariate analyses were included in a logistic model and then those symptoms that were not significant statistically ($P > 0.1$) were dropped from the model in a backward stepwise manner. Coefficients of each symptom remaining in the model were summed to obtain a score for each subject i.e. Score = $b_{1\times1}+b_{2\times2}+\ldots$, where b_ix_i are the log OR b_i of symptoms x_i in the model. A cut-off score was identified for each cause of death (included 16 primary causes of adult death) that gave the estimated number of deaths closest to the true number of cause-specific deaths, such that the sensitivity was at least 50%.

We used the same train and test datasets used by Quigley *et al.* for training and testing an ANN. The data were ported to Microsoft Excel™ and analysed using NeuroSolutions 3.0™ (Lefebvre WC. NeuroSolution Version 3.020, Neurodimension

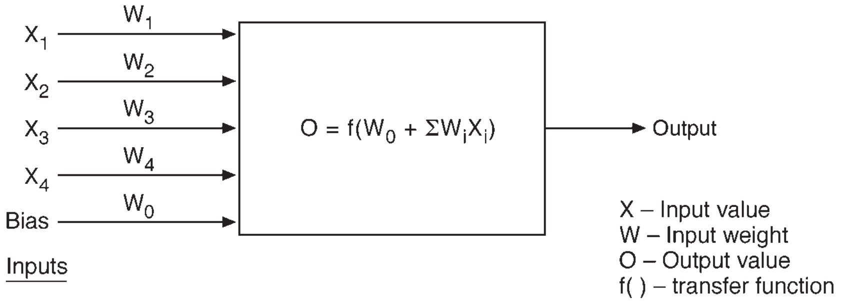

Figure 1 Schematic representation of a single node in a neural network

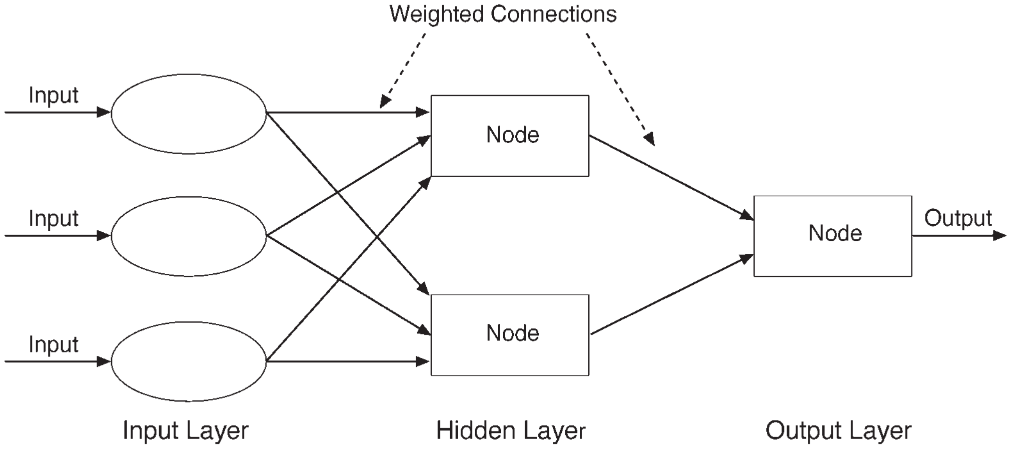

Figure 2 Schematic representation of multi-layer perceptron

Inc.1994. [www.nd.com]). All models were multi-layer perceptrons with a single hidden layer and trained with static backpropogation. The number of nodes in the hidden layer were varied according to the number of inputs and network performance. A learning rate of 0.7 was used throughout with the momentum learning rule. A sigmoid activation function was used for all processing elements.

Model inputs were based on those used in the logistic regression study, with further variables added to improve discrimination in instances when they improved the model performance. Sensitivity analysis provided the basis for evaluating the role of the inputs in the models.

For each diagnosis, the first 100 records of the training subset were used in the first training run of each model as a cross-validation set to determine the optimal number of hidden nodes and the training point at which the cross-validation mean squared error reached a trough. Thereafter the full training set was used to train the network to this point.

The output weights were then adjusted by a variable factor until the CSMF was as close as possible to 100% of the expected value in testing runs on the training set. At this point the network was tested on the unseen data in the test subset.

Weighted (by number of deaths) averages for sensitivity and specificity were calculated for each method. A summary measure for CSMF was calculated for each method by summing the absolute difference in observed and estimated number of cases for each cause of death, dividing by the total number of deaths, and converting to a percentage.

Results

Table 1 shows the comparison of validity of the logistic regression models versus the ANN models for estimating CSMF by comparing estimated with observed number of cases as well as sensitivity and specificity.

The CSMF was estimated to within 10% of the true value in 8 out of 16 classes (causes of death) by the ANN. In a further six classes it was estimated to within 25% of the true value. In the remaining two classes the CSMF was extremely low (tetanus and rabies). The summary measure for CSMF favours those methods that are more accurate on the more frequently occurring classes and may mask poor performance on rare causes of death. In this measure however, calculated from the absolute number of over- or under-diagnosed cases, the neural network method performed better than logistic regression models (average error 11.27% versus 31.27%), and compared well with physician review (average error of 12.84%). In the assessment of chance agreement between ANN and gold-standard diagnoses,

Table 1 Comparison of performance of physician review, logistic regression and neural network models

	Combined dataset				**Test dataset**							
		Physician review				Logistic regression			Neural network			
Causes of death	Observed	Est.[a]	Sens.[b]	Spec.[c]	Observed	Est.[a]	Sens.[b]	Spec.[c]	Est.[a]	Sens.[b]	Spec.[c]	Kappa
TB-AIDS[d]	148	**151**	76	94	71	**78**	65	90	**77**	66	90	0.55
Malaria	85	**80**	33	93	42	89	48	80	**44**	29	91	0.19
Meningitis	66	**60**	59	97	32	**33**	53	95	**34**	56	95	0.50
CVS[e] disorders	65	*57*	48	96	32	**32**	34	94	**30**	47	97	0.44
Acute abdominal conditions	55	*65*	69	97	27	18	37	98	**29**	48	96	0.42
Diarrhoeal diseases	51	*58*	61	96	25	*30*	48	95	*31*	60	96	0.50
Direct maternal causes	50	*58*	82	98	25	18	52	97	*19*	48	98	0.52
Neoplasms	34	**36**	50	98	16	24	19	94	*14*	6	96	0.03
Injuries	33	**36**	97	99.5	15	**16**	80	99	**16**	80	99	0.76
Hepatitis	32	14	34	99.6	16	10	0	97	*19*	6	95	0.01
Chronic liver disease	25	*29*	40	98	12	*14*	8	97	*15*	25	97	0.19
Anaemia	24	12	33	99.5	12	25	25	94	**11**	17	98	0.15
Pneumonia	23	*19*	39	98.7	11	14	27	97	*13*	27	97	0.23
Renal disorders	21	**22**	38	98	10	**10**	10	98	**10**	10	98	0.08
Tetanus	13	*10*	77	100	6	*5*	17	99	3	50	100	0.66
Rabies	7	**7**	86	99.9	3	4	100	99.7	4	100	0.86	
Weighted average[f]			58.95	96.39			43.98	93.01		45.32	94.72	
Summary CSMF[g] accuracy[h]		12.84				31.27			11.83			

[a] Estimated.

[b] Sensitivity.

[c] Specificity.

[d] Tuberculosis-autoimmune deficiency syndrome.

[e] Cardiovascular system.

[f] ∑ Sensitivity or specificity of each class × number of cases in each class/total number of cases.

[g] Cause-specific mortality fraction.

[h] ∑ Difference between observed and estimated cases in each class/total number of cases.

Figures in Bold: Estimated CSMF is within 10% of expected value.

Figures in italics: Estimated CSMF is within 25% of expected value.

the kappa value was ⩾0.5 for the following classes: rabies (0.86), injuries (0.76), tetanus (0.66), tuberculosis and AIDS (0.55), direct maternal causes (0.52), meningitis (0.50), and diarrhoea (0.50).

There was a trade-off between specificity and sensitivity, and in some instances the neural network performed better than other techniques in one at the expense of the other. Compared to logistic regression, the networks performed better in both parameters for tuberculosis and AIDS, meningitis, cardiovascular disorders, diarrhoea, and tetanus. They produced a lower sensitivity for malaria (compared with logistic regression), but a higher specificity. The overall and disease-specific sensitivities and specificities compared favourably with logistic regression, but did not match the performance of physician review.

Discussion

Accuracy of CSMF estimates

One of the most significant findings of this analysis is the relative accuracy in assessing the fraction of deaths that are due to specific causes, especially for the more frequently occurring classes. The accuracy in this estimate does not always correlate with the reliability estimated by the kappa statistic. Care was taken to find a weighting for the output that would lead to a correct CSMF in the training set. The choice of this weight is analogous to selecting the minimum total score at which a case is defined in the logistic regression models. This then led to surprisingly good estimates in the testing set. It is a feature of the train and test subsets however that the number of members in each class is similar. Manipulating either subset so that the CSMF differed, by randomly removing or adding records of the class in question, did not alter the accuracy of the CSMF estimates if the number of training examples for the class was not decreased in the training subset. With less frequently occurring classes such as pneumonia, decreasing the number of training examples in the training set, reduced the accuracy of the CSMF estimate. This is essentially an issue of generalisation, and it is to be expected that networks that are trained with fewer examples are less likely to be generalizable. It is suggested that it is for this reason that the CSMF estimates for the five most frequently occurring classes are all within 10% of the expected values. It would be expected furthermore that if the datasets were larger, that the generalizability of the CSMF estimates for the less frequently occurring classes would improve.

At the stage of data analysis the question can be asked as to whether or not there is an output level above which class membership is reasonably certain, and below which misclassification is more likely to occur. Looking at the tuberculosis-AIDS model (n = 71), as well as the meningitis model (n = 32), and ranking the top 20 test outputs in descending order by value (reflecting the certainty of the classification), 13/20 of these outputs correctly predict the class membership in both instances. The sensitivities for the models overall were 66% and 56% respectively. The implication is that without a gold standard result for comparison, it would be difficult to delineate the true positives from the false positives even in the least equivocal outputs. This is in keeping with observations that different data-derived methods arrive at their estimates differently. One study to predict an acute abdomen diagnosis from surgical admission records demonstrated that data-derived methods with similar overall performance correlated poorly as to which of the records they were correctly predicting.[7]

Mechanisms of improved performance

A single layer neural network (i.e. a network with only inputs, and one processing element) is isomorphic with logistic regression. A network with no hidden nodes produced almost identical results when comparing the input weights to the log(OR) for the four inputs used in the regression model to predict malaria as the cause of death. In those instances where the performance of logistic regression and neural network models differ, it is of interest as to know the mechanisms by which improvements are made. The results from this study indicate that the differences in performance of the neural networks are achieved both by improved fitting of those variables already known to be significantly predictive of class membership, through the modelling of interaction between them, and by additional discriminating power conferred by variables that are not significantly predictive on their own

The first mechanism was borne out in one of the meningitis models in which the exact same inputs used in the logistic regression model were used in the neural network model with an improvement in performance. Exploring the sensitivity analysis for cardiovascular deaths (Table 2), the network outputs are surprisingly sensitive to the absence of a tuberculosis history, which was not strongly predictive by itself. Age above 45 years old was the seventh most predictive input in the regression model, whereas it was the input to which the neural network model was second most sensitive. In the case of meningitis, presence of continuous fever was more important in the

Table 2 Comparison of the most important inputs for two data-derived models for assigning cardiovascular deaths

Logistic regression model		Neural network model	
Rank[a]	Input	Rank[b]	Input
1	Puffiness of face	1	Puffiness of face
2	Cough 3–14 days	2	Age ⩾45years
3	Abdominal distension 8–30 days	3	No weight loss
4	No weight loss	4	Abdominal distension 8–30 days
5	No jaundice	5	History of hypertension
6	History of hypertension	6	No history of tuberculosis
7	Age ⩾45years	7	No jaundice
		8	No recent surgery
		9	Pallor
		10	No stiff neck
		11	Cough >3 weeks
		12	Shortness of breath
		13	No chronic diarrhoea
		14	No productive cough
		15	Chest pain
		16	Wheeze
		17	No Continuous fever

[a] As determined by the log (odds ratio) for each input.

[b] As determined by sensitivity analysis in which the standard deviation of the output response as the input is varied, is divided by the standard deviation of the input.

regression model, whilst presence or absence of recent surgery and abdominal distension were more significant in the ANN model (Table 3). The network has mapped relationships between the inputs that were not predicted by the regression model.

Effect of size of dataset

Both data-derived methods stand to benefit from more training examples. In the regression models, some inputs not currently utilized may yield significant associations with outputs when larger datasets are used. With enough nodes and training time, it was possible in the course of this analysis to train a neural network to completely map the training set with 100% sensitivity and specificity. However, this level of sensitivity and specificity was not reproduced when these models were tested in the test dataset. What it did demonstrate is the ability of the method to map complex functions. The key point is one of generalizability. In the models presented above, training was stopped and the nodes limited to ensure that the generalizability was not compromised. With more training examples, it is likely that the networks would develop a better understanding of the relationships between inputs and outputs before over-training occurs. Arguably, the neural network models would stand to improve performance more than the regression models should larger training sets be available. However, further training may not achieve algorithms of sufficiently high sensitivity and specificity to obviate the need for algorithms with particular operating characteristics suitable for use in specific environments.

Physician review

Only 78% of the reference diagnoses were confirmed by laboratory tests. Since 22% of the reference diagnoses were based on hospital physicians' clinical judgement, it is not surprising that physician review of VA performed better than the other methods. Nevertheless, physician review remains the optimal method of analysis, as far as overall performance is concerned, for gathering cause-specific mortality data as good as the data produced by routine health information systems.[1] The technique by which physicians in this study came to their classification differed considerably, as they made extensive use of the open section of the questionnaire from which information was not coded for analysis by the other techniques. Interestingly though, other methods are able to come close if the CSMF is used as the outcome of choice, as indeed it often is. Thus ANN or logistic regression models based algorithms have the potential for substituting physician review of VA.

Table 3 Comparison of the most important inputs for two data-derived models for assigning death due to meningitis

Logistic regression model		Neural network model	
Rank[a]	Input	Rank[b]	Input
1	Stiff neck	1	Stiff neck
2	No cough	2	Cough
3	No pallor	3	Recent surgery
4	Continuous fever	4	Pallor
5	Stiff body	5	Abdominal distension
		6	Injury/accident
		7	Severe loss of weight
		8	Body stiffness
		9	Tuberculosis

[a] As determined by the log (odds ratio) for each input.

[b] As determined by sensitivity analysis in which the standard deviation of the output response as the input is varied, is divided by the standard deviation of the input.

Limitations of the technique

At various points we have alluded to some of the difficulties and limitations of using neural networks for the analysis. These are summarized in Table 4.

Even with sensitivity analysis, we had no way of working out which were going to be the most important inputs prior to creating a model and conducting a sensitivity analysis on it. There is some correlation with linearly predictive inputs that helps in the initial stages.

Determining the weighting for the output for providing the optimum estimate of the CSMF was time-consuming. The software provides an option for prioritizing sensitivity over specificity, but no way of balancing the number of false positives and false negatives that would give an accurate CSMF estimate.

Designing the optimal network topology requires building numerous networks in search of the one with the lowest least mean squared error. The number of hidden nodes, inputs and training time all affect the performance of the network. Whilst training is relatively quick compared to the many hours it took to train ANN in the early days of their development, it is still time-consuming to build and train multiple networks for each model.

Cross-validation to prevent over-training required compromising the number of training examples to allow for a cross-validation dataset.

Sensitivity and specificity of the ANN algorithms were not high enough to be generalizable to a variety of settings. Furthermore, the accuracy of individual and summary estimates of CSMF obtained in this study could be due to the similarity in the CSMF between the training and test datasets. Thus large datasets from a variety of settings are needed to identify optimal algorithms for each site with different distributions of causes of death.

Conclusions

Classification software based on neural network simulations is an accessible tool which can be applied to VA data potentially outperforming other the data-derived techniques already studied for this purpose. As with other data-derived techniques, over-fitting to the training data leading to a compromise in the generalizability of the models is a potential limitation of ANN. Increasing the number of training examples is likely to improve performance of neural networks for VA. However, ANN algorithms with particular operating characteristics would be site-specific. Thus optimal algorithms need to be identified for use in a variety of settings.

Table 4 Limitations of the artificial neural network technique

Selecting inputs is not straightforward
Prioritizing cause-specific mortality fraction over sensitivity or specificity is a manual process
Designing optimal networks for each cause of death is time-consuming
Sensitivity and specificity may be not high enough for the algorithms to be generalizable to a variety of settings

KEY MESSAGES

- Artifical neural networks have potential for classifying causes of death from verbal autopsies.
- Large datasets are needed to train neural networks and for validating their performance.
- Generalizability of neural network models to various settings needs further evaluation.

References

[1] Chandramohan D, Maude H, Rodrigues L, Hayes R. Verbal autopsies for adult deaths: their development and validation in a multicentre study. *Trop Med Int Health* 1998;**3:**436–46.

[2] Snow RW, Armstrong ARM, Forster D *et al*. Childhood deaths in Africa: uses and limitations of verbal autopsies. *Lancet* 1992;**340:**351–55.

[3] Todd JE, De Francisco A, O'Dempsey TJD, Greenwood BM. The limitations of verbal autopsy in a malaria-endemic region. *Ann Trop Paediatr* 1994;**14:**31–36.

[4] Quigley M, Chandramohan D, Rodrigues L. Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. *Int J Epidemiol* 1999;**28:**1081–87.

[5] Jefferson MF, Pendleton N, Lucas SB, Horan MA. Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. *Cancer* 1997;**79:**1338–42.

[6] Duh MS, Walker AM, Pagano M, Kronlund K. Epidemiological interpretation of artificial neural networks. *Am J Epidemiol* 1998;**147:** 1112–22.

[7] Schwartz S, *et al*. Connectionist, rule-based and Bayesian diagnostic decision aids: an empirical comparison. In: Hand DJ (ed.). *Artificial Intelligence Frontiers in Statistics*. London: Chapman and Hall, 1993, pp.264–77.