
A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer

Tony Plate

BC Cancer Agency
601 West 10th Ave, Epidemiology
Vancouver BC Canada V5Z 1L3
tap@comp.vuw.ac.nz

Pierre Band

BC Cancer Agency
601 West 10th Ave, Epidemiology
Vancouver BC Canada V5Z 1L3

Joel Bert

Dept of Chemical Engineering
University of British Columbia
2216 Main Mall
Vancouver BC Canada V6T 1Z4

John Grace

Dept of Chemical Engineering
University of British Columbia
2216 Main Mall
Vancouver BC Canada V6T 1Z4

Abstract

Epidemiological data is traditionally analyzed with very simple techniques. Flexible models, such as neural networks, have the potential to discover unanticipated features in the data. However, to be useful, flexible models must have effective control on overfitting. This paper reports on a comparative study of the predictive quality of neural networks and other flexible models applied to real and artificial epidemiological data. The results suggest that there are no major unanticipated complex features in the real data, and also demonstrate that MacKay's [1995] Bayesian neural network methodology provides effective control on overfitting while retaining the ability to discover complex features in the artificial data.

1 Introduction

Traditionally, very simple statistical techniques are used in the analysis of epidemiological studies. The predominant technique is logistic regression, in which the effects of predictors are linear (or categorical) and additive on the log-odds scale. An important virtue of logistic regression is that the relationships identified in the

data can be interpreted and explained in simple terms, such as "the odds of developing lung cancer for males who smoke between 20 and 29 cigarettes per day are increased by a factor of 11.5 over males who do not smoke". However, because of their simplicity, it is difficult to use these models to discover unanticipated complex relationships, i.e., non-linearities in the effect of a predictor or interactions between predictors. Interactions and non-linearities can of course be introduced into logistic regressions, but must be pre-specified, which tends to be impractical unless there are only a few variables or there are a priori reasons to test for particular effects.

Neural networks have the potential to automatically discover complex relationships. There has been much interest in using neural networks in biomedical applications; witness the recent series of articles in *The Lancet*, e.g., Wyatt [1995] and Baxt [1995]. However, there are not yet sufficient comparisons or theory to come to firm conclusions about the utility of neural networks in biomedical data analysis. To date, comparison studies, e.g, those by Michie, Spiegelhalter, and Taylor [1994], Burke, Rosen, and Goodman [1995], and Lippmann, Lee, and Shahian [1995], have had mixed results, and Jefferson et al's [1995] complaint that many "successful" applications of neural networks are not compared against standard techniques appears to be justified. The intent of this paper is to contribute to the body of useful comparisons by reporting a study of various neural-network and statistical modeling techniques applied to an epidemiological data analysis problem.

2 The data

The original data set consisted of information on 15,463 subjects from a study conducted by the Division of Epidemiology and Cancer Prevention at the BC Cancer Agency. In this study, detailed questionnaire reported personal information, lifetime tobacco and alcohol use, and lifetime employment history for each subject. The subjects were cancer patients in BC with diagnosis dates between 1983 and 1989, as ascertained by the population-based registry at the BC Cancer Agency. Six different tobacco and alcohol habits were included: cigarette (C), cigar (G), and pipe (P) smoking, and beer (B), wine (W), and spirit drinking (S). The models reported in this paper used up to 27 predictor variables: age at first diagnosis (AGE), and 26 variables related to alcohol and tobacco consumption. These included four variables for each habit: total years of consumption (CYR etc), consumption per day or week (CDAY, BWK etc), years since quitting (CYQUIT etc), and a binary variable indicating any indulgence (CSMOKE, BDRINK etc). The remaining two binary variables indicated whether the subject ever smoked tobacco or drank alcohol. All the binary variables were non-linear (threshold) transforms of the other variables. Variables not applicable to a particular subject were zero, e.g., number of years of smoking for a non-smoker, or years since quitting for a smoker who did not quit.

Of the 15,463 records, 5901 had missing information in some of the fields related to tobacco or alcohol use. These were not used, as there are no simple methods for dealing with missing data in neural networks. Of the 9,562 complete records, a randomly selected 3,195 were set aside for testing, leaving 6,367 complete records to be used in the modeling experiments.

There were 28 binary outcomes: the 28 sites at which a subject could have cancer (subjects had cancers at up to 3 different sites). The number of cases for each site varied, e.g., for LUNGSQ (Lung Squamous) there were 694 cases among the complete records, for ORAL (Oral Cavity and Pharynx) 306, and for MEL (Melanoma) 464.

All sites were modeled individually using carefully selected subjects as controls. This is common practice in cancer epidemiology studies, due to the difficulty of collecting an unbiased sample of non-cancer subjects for controls. Subjects with

cancers at a site suspected of being related to tobacco usage were not used as controls. This eliminated subjects with any sites other than COLON, RECTUM, MEL (Melanoma), NMSK (Non-melanoma skin), PROS (Prostate), NHL (Non-Hodgkin's lymphoma), and MMY (Multiple-Myeloma), and resulted in between 2959 and 3694 controls for each site. For example, the model for LUNGSQ (lung squamous cell) cancer was fitted using subjects with LUNGSQ as the positive outcomes (694 cases), and subjects all of whose sites were among COLON, RECTUM, MEL, NMSK, PROS, NHL, or MMY as negative outcomes (3694 controls).

3 Statistical methods

A number of different types of statistical methods were used to model the data. These ranged from the non-flexible (logistic regression) through partially flexible (Generalized Additive Models or GAMs) to completely flexible (classification trees and neural networks). Each site was modeled independently, using the log likelihood of the data under the binomial distribution as the fitting criterion. All of the modeling, except for the neural networks and ridge regression, was done using the the S-plus statistical software package [StatSci 1995].

For several methods, we used Breiman's [1996] *bagging* technique to control overfitting. To "bag" a model, one fits a set of models independently on bootstrap samples. The bagged prediction is then the average of the predictions of the models in the set. Breiman suggests that bagging will give superior predictions for unstable models (such as stepwise selection, pruned trees, and neural networks).

Preliminary analysis revealed that the predictive power of non-flexible models could be improved by including non-linear transforms of some variables, namely AGESQ and the binary indicator variables SMOKE, DRINK, CSMOKE, etc. Flexible models should be able to discover useful non-linear transforms for themselves and so these derived variables were not included in the flexible models. In order to allow comparisons to test this, one of non-flexible models (ONLYLIN-STEP) also did not use any of these derived variables.

Null model: (NULL) The predictions of the null model are just the frequency of the outcome in the training set.

Logistic regression: The FULL model used the full set of predictor variables, including a quadratic term for age: AGESQ.

Stepwise logistic regression: A number of stepwise regressions were fitted, differing in the set of variables considered. Outcome-balanced 10-fold cross validation was used to select the model size giving best generalization. The models were as follows: AGE-STEP (AGE and AGESQ); CYR-AGE-STEP (CYR, AGE and AGESQ); ALC-CYR-AGE-STEP (all alcohol variables, CYR, AGE and AGESQ); FULL-STEP (all variables including AGESQ); and ONLYLIN-STEP (all variables except for the derived binary indicator variables SMOKE, CSMOKE, etc, and only a linear AGE term).

Ridge regression: (RIDGE) Ridge regression penalizes a logistic regression model by the sum of the squared parameter values in order to control overfitting. The evidence framework [MacKay 1995] was used to select seven shrinkage parameters: one for each of the six habits, and one for SMOKE, DRINK, AGE and AGESQ.

Generalized Additive Models: GAMs [Hastie and Tibshirani 1990] fit a smoothing spline to each parameter. GAMs can model non-linearities, but not interactions. A stepwise procedure was used to select the degree (0,1,2, or 4) of the smoothing spline for each parameter. The procedure started with a model having a smoothing spline of degree 2 for each parameter, and stopped when the AIC statistic could

not reduced any further. Two stepwise GAM models were fitted: GAM-FULL used the full set of variables, while GAM-CIG used the cigarette variables and AGE.

Classification trees: [Breiman et al. 1984] The same cross-validation procedure as used with stepwise regression was used to select the best size for TREE, using the implementation in S-plus, and the function `shrink.tree()` for pruning. A bagged version with 50 replications, TREE-BAGGED, was also used. After constructing a tree for the data in a replication, it was pruned to perform optimally on the training data not included in that replication.

Ordinary neural networks: The neural network models had a single hidden layer of tanh functions and a small weight penalty (0.01) to prevent parameters going to infinity. A conjugate-gradient procedure was used to optimize weights. For the NN-ORD-H2 model, which had no control on complexity, a network with two hidden units was trained three times from different small random starting weights. Of these three, the one with best performance on the training data was selected as “the model”. The NN-ORD-HCV used common method for controlling overfitting in neural networks: 10-fold CV for selecting the optimal number of hidden units. Three random starting points for each partition were used calculate the average generalization error for networks with one, two and three hidden units Three networks with the best number of hidden units were trained on the entire set of training data, and the network having the lowest training error was chosen.

Bagged neural networks with early stopping: Bagging and early stopping (terminating training before reaching a minimum on training set error in order to prevent overfitting) work naturally together. The training examples omitted from each bootstrap replication provide a validation set to decide when to stop, and with early stopping, training is fast enough to make bagging practical. 100 networks with two hidden units were trained on separate bootstrap replications, and the best 50 (by their performance on the omitted examples) were included in the final bagged model, NN-ESTOP-BAGGED. For comparison purposes, the mean individual performance of these early-stopped networks is reported as NN-ESTOP-AVG.

Neural networks with Bayesian regularization: MacKay’s [1995] Bayesian evidence framework was used to control overfitting in neural networks. Three random starts for networks with 1, 2, 3 or 4 hidden units and three different sets of regularization (penalty) parameters were used, giving a total of 36 networks for each site. The three possibilities for regularization parameters were: (a) three penalty parameters – one for each of input to hidden, bias to hidden, and hidden to output; (b) partial Automatic Relevance Determination (ARD) [MacKay 1995] with seven penalty parameters controlling the input to hidden weights – one for each habit and one for AGE; and (c) full ARD, with one penalty parameter for each of the 19 inputs. The “evidence” for each network was evaluated and the best 18 networks were selected for the equally-weighted committee model NN-BAYES-CMTT. NN-BAYES-BEST was the single network with the maximum evidence.

4 Results and Discussion

Models were compared based on their performance on the held-out test data, so as to avoid overfitting bias in evaluation. While there are several ways to measure performance, e.g., 0-1 classification error, or area under the ROC curve (as in Burke, Rosen and Goodman [1995]), we used the test-set deviance as it seems appropriate to compare models using the same criterion as was used for fitting. Reporting performance is complicated by the fact that there were 28 different modeling tasks (i.e., sites), and some models did better on some sites and worse on others. We report some overall performance figures and some pairwise comparisons of models.

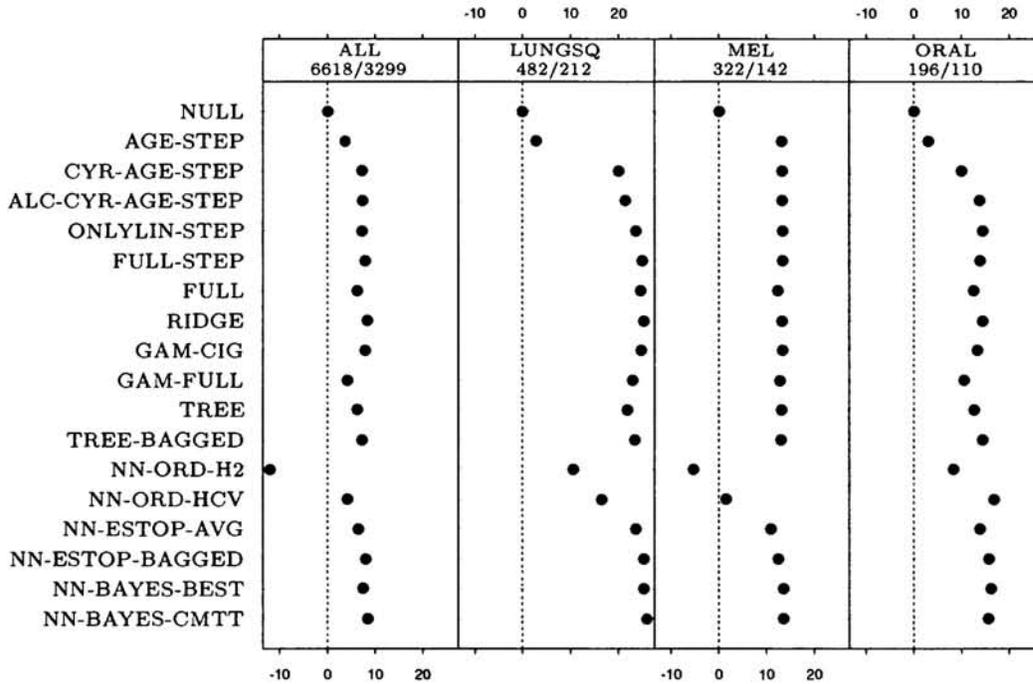


Figure 1: Percent improvement in deviance on test data over the null model.

Figure 1 shows aggregate deviances across sites (i.e., the sum of the test deviance for one model over the 28 sites) and deviances for selected sites. The horizontal scale in each column indicates the percentage reduction in deviance over the null model. Zero percent (the dotted line) is the same performance as the null model, and 100% would be perfect predictions. Numbers below the column labels are the number of positive outcomes in the training and test sets, respectively. The best predictions for LUNGSQ can reduce the null deviance by just over 25%. It is interesting to note that much of the information is contained in AGE and CYR: The CYR-AGE-STEP model achieved a 7.1% reduction in overall deviance, while the maximum reduction (achieved by NN-BAYES-CMTT) was only 8.3%.

There is no single threshold at which differences in test-set deviance are “significant”, because of strong correlations between predictions of different models. However, the general patterns of superiority apparent in Figure 1 were repeated across the other sites, and various other tests indicate they are reliable indicators of general performance. For example, the best five models, both in terms of aggregate deviance across all sites and median rank of performance on individual sites, were, in order NN-BAYES-CMTT, RIDGE, NN-ESTOP-BAGGED, GAM-CIG, and FULL-STEP. The ONLYLIN-STEP model ranked sixth in median rank, and tenth in aggregate deviance.

Although the differences between the best flexible models and the logistic models were slight, they were consistent. For example, NN-BAYES-CMTT did better than FULL-STEP on 21 sites, and better than ONLYLIN-STEP on 23 sites, while FULL-STEP drew with ONLYLIN-STEP on 14 sites and did better on 9. If the models had no effective difference, there was only a 1.25% chance of one model doing better than the other 21 or more times out of 28. Individual measures of performance were also consistent with these findings. For example, for LUNGSQ a bootstrap test of test-set deviance revealed that the predictions of NN-BAYES-CMTT were on average better than those of ONLYLIN-STEP in 99.82% of resampled test sets (out of 10,000), while the predictions of NN-BAYES-CMTT beat FULL-STEP in 93.75% of replications

and FULL-STEP beat ONLYLIN-STEP in 98.48% of replications.

These results demonstrate that good control on overfitting is essential for this task. Ordinary neural networks with no control on overfitting do worse than guessing (i.e., the null model). Even when the number of hidden units is chosen by cross-validation, the performance is still worse than a simple two-variable stepwise logistic regression (CYR-AGE-STEP). The inadequacy of the simple AIC-based stepwise procedure for choosing the complexity of GAMs is illustrated by the poor performance of the GAM-FULL model (the more restricted GAM-CIG model does quite well).

The effective methods for controlling overfitting were bagging and Bayesian regularization. Bagging improved the performance of trees and early-stopped neural networks to good levels. Bayesian regularization worked very well with neural networks and with ridge regression. Furthermore, examination of the performance of individual networks indicates that networks with fine-grained ARD were frequently superior to those with coarser control on regularization.

5 Artificial sites with complex relationships

The very minor improvement achieved by neural networks and trees over logistic models provokes the following question: are complex relationships really relatively unimportant in this data, or is the strong control on overfitting preventing identification of complex relationships? In order to answer this question, we created six artificial “sites” for the subjects. These were designed to have very similar properties to the real sites, while possessing non-linear effects and interactions.

The risk models for the artificial sites possessed a underlying trend equal to half that of a good logistic model for LUNGSQ, and one of three more complex effects: FREQ, a frequent non-linear (threshold) effect ($BWK > 1$) affecting 4,334 of the 9,562 subjects; RARE, a rare threshold effect ($BWK > 10$), affecting 1,550 subjects; and INTER, an interaction ($BYR \cdot GYR$) affecting 482 subjects. For three of the artificial sites the complex effect was weak (LO), and for the other three it was strong (HI). For each subject and each artificial site, a random choice as to whether that subject was a positive case for that site was made, based on probability given by the model for the artificial site. Models were fitted to these sites in the same way as to other sites and only subjects without cancer at a smoking related site were used as controls.

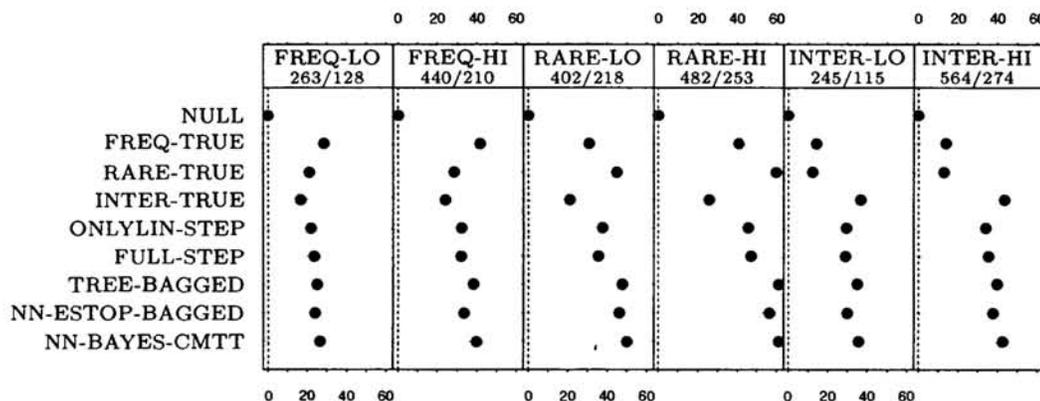


Figure 2: Percent improvement in deviance on test data for the artificial sites.

For comparison purposes, logistic models containing the true set of variables, including non-linearities and interactions, were fitted to the artificial data. For example, the model RARE-TRUE contained the continuous variables AGE, AGESQ, CDAY, CYR, and CYQUIT, and the binary variables SMOKE and $BWK > 10$.

Figure 2 shows performance on the artificial data. The neural networks and bagged trees were very effective at detecting non-linearities and interactions. Their performance was at the same level as the appropriate true models, while the performance of simple models lacking the ability to fit the complexities (e.g., FULL-STEP) was considerably worse.

6 Conclusions

For predicting the risk of cancer in our data, neural networks with Bayesian estimation of regularization parameters to control overfitting performed consistently but only slightly better than logistic regression models. This appeared to be due to the lack of complex relationships in the data: on artificial data with complex relationships they performed markedly better than logistic models. Good control of overfitting is essential for this task, as shown by the poor performance of neural networks with the number of hidden units chosen by cross-validation.

Given their ability to not overfit while still identifying complex relationships we expect that neural networks could prove useful in epidemiological data-analysis by providing a method for checking that a simple statistical model is not missing important complex relationships.

Acknowledgments

This research was funded by grants from the Workers Compensation Board of British Columbia, NSERC, and IRIS, and conducted at the BC Cancer Agency.

References

- Baxt, W. G. 1995. Application of artificial neural networks to clinical medicine. *The Lancet*, 346:1135–1138.
- Breiman, L. 1996. Bagging predictors. *Machine Learning*, 26(2):123–140.
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, CA.
- Burke, H., Rosen, D., and Goodman, P. 1995. Comparing the prediction accuracy of artificial neural networks and other statistical methods for breast cancer survival. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, *Advances in Neural Information Processing Systems 7*, pages 1063–1067, Cambridge, MA. MIT Press.
- Hastie, T. J. and Tibshirani, R. J. 1990. *Generalized additive models*. Chapman and Hall, London.
- Jefferson, M. F., Pendleton, N., Lucas, S., and Horan, M. A. 1995. Neural networks (letter). *The Lancet*, 346:1712.
- Lippmann, R., Lee, Y., and Shahian, D. 1995. Predicting the risk of complications in coronary artery bypass operations using neural networks. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, *Advances in Neural Information Processing Systems 7*, pages 1055–1062, Cambridge, MA. MIT Press.
- MacKay, D. J. C. 1995. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6:469–505.
- Michie, D., Spiegelhalter, D., and Taylor, C. 1994. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Hertfordshire, UK.
- StatSci 1995. *S-Plus Guide to Statistical and Mathematical Analyses, Version 3.3*. StatSci, a division of MathSoft, Inc, Seattle.
- Wyatt, J. 1995. Nervous about artificial neural networks? (commentary). *The Lancet*, 346:1175–1177.