Human and Ecological Risk Assessment

ISSN: 1080-7039 (Print) 1549-7860 (Online) Journal homepage: https://www.tandfonline.com/loi/bher20

Identifying Populations at High Risk for Occupational Back Injury with Neutral Networks

Douglas P. Landsittel , Lytt I. Gardner & Vincent C. Arena

To cite this article: Douglas P. Landsittel , Lytt I. Gardner & Vincent C. Arena (1998) Identifying Populations at High Risk for Occupational Back Injury with Neutral Networks, Human and Ecological Risk Assessment, 4:6, 1337-1352, DOI: 10.1080/10807039891284703

To link to this article: https://doi.org/10.1080/10807039891284703

Published online: 03 Jun 2010.

Human and Ecological Risk Assessment: Vol. 4, No. 6, pp. 1337-1352 (1998)

Identifying Populations at High Risk for Occupational Back Injury with Neutral Networks

Douglas P. Landsittel,[1],* Lytt I. Gardner,[2], and Vincent C. Arena[3],*****
[1]National Institute for Occupational Safety and Health, Morgantown, WV;
[2]National Institute for Occupational Safety and Health, Morgantown, WV;
[3]University of Pittsburgh, Pittsburgh, PA

ABSTRACT

For this study a simulation is conducted to investigate the accuracy of neural networks and logistic regression in identifying populations at high risk for occupational back injury. In contrast to most standard regression techniques, neural networks do not rely on linearity or explicitly specifying the nature of the association. Because the underlying relationships between work exposures, personal risk factors, and injury are often not well defined, neural networks may prove useful for injury risk assessment. Accuracy was assessed by comparing the injury status to the predicted level of risk in each worker. In simulations of a non-linear association, workers (used in the training data) were correctly classified 85% of the time with neural networks, 74% of the time with the main effects logistic model, and 79% of the time with the fully-specified logistic model. Using the test data, however, workers were correctly classified 67% of the time with neural networks, and 71% and 69% of the time with the main effects and fully-specified logistic models, respectively. Simulations of a null association indicated that neural networks may be more likely to overfit random associations. These findings provide a valuable guide concerning statistical methodology for identifying high-risk worker populations.

Key Words: classification, logistic regression, simulation study.

* Corresponding author. National Institute for Occupational Safety and Health, 1095 Willowdale Road, M/S P1133, Morgantown, WV 26505; Tel: (304) 285-6075; Fax: (304) 285-6047

** National Institute for Occupational Safety and Health, 1095 Willowdale Road, M/S P1133, Morgantown, WV 26505; Tel: (304) 285-6075; Fax: (304) 285-6047

*** Department of Biostatistics, University of Pittsburgh, 130 DeSoto Street, Pittsburgh, PA 15261; Tel: (412) 624-3023; Fax: (412) 624-2183

INTRODUCTION

An important aspect of occupational injury research is the assessment of an individual worker's risk for injury (Courtney *et al.*, 1997). By classifying each worker into low- or high-risk categories we can identify which segment of the population is at high risk for occupational back injury. Identification of a high-risk population allows researchers to better focus and evaluate interventions or treatments. Classification results from neural networks are compared to classification results from logistic regression to assess each methods' accuracy in identifying high-risk populations.

The motivation for utilizing neural networks in this setting is to address the analysis of non-linear associations in the data. The neural network model does not make any assumptions about the nature of an association between the outcome and predictors (Stern, 1996). If the exact nature of these relationships are known then the probability of injury can be explicitly modeled with logistic regression (Hosmer and Lemeshow, 1989) using appropriate categories or transformations. Otherwise, if we cannot explicitly specify the general nature of the non-linear association (such as quadratic or cubic), extensive exploratory analysis becomes necessary. Due to practical limitations in cases where a very small percentage of the population is injured, adequate data may not exist for thoroughly investigating the underlying structure of the data. Neural networks offer another approach to analyzing non-linear associations when we cannot adequately define the nature of the associations based on prior knowledge (Ripley, 1993). Other methods, such as generalized additive models, offer additional approaches which may be more easily interpreted. Comparisons between neural nets and such methods are not explored in this study. Methods for interpreting neural network parameters are less developed with neural networks than with standard methods (Lippmann and Shahian, 1997), thus providing further motivation to avoid using neural nets for statistical inference / estimation if the underlying data structure can be adequately described. Very few publications have thoroughly researched parameter interpretation with neural nets.

Neural networks should only be thought of as an exploratory technique in that the nature of the association between predictors and outcome is implicitly determined. In contrast to standard regression methods, further knowledge about this relationship, other than which variables to include in the model, is not required (or even useful) for neural network analysis. The network transforms the data to find the optimal classification of (for instance) cases and controls. Significance testing and estimation of summary measures is possible, although more difficult, with neural networks (Lippmann and Shahian, 1997; Landsittel, 1997). Neural networks are typically implemented for prediction, especially when the underlying structure of the data is very complex and / or unknown.

Numerous statistical techniques have been implemented for the purposes of classification, or identification of high-risk populations. Logistic regression, which uses maximum likelihood methods to fit the data to a linear function (in the logit scale) of the predictors and interactions, often serves as the standard statistical method for classification (Tu and Guerriere, 1993; Ripley, 1994; Tu, 1996, Duh *et al.*, 1998b). Other methods, such as probit analysis, discriminant analysis (Anderson, 1984) and classification and regression trees (Breiman, 1984), utilize different criteria or different assumptions to determine the optimal classification model. Modern regression techniques, such as projection pursuit regression (Friedman, 1987; Jones and Sibson, 1987) and multivariate adaptive regression splines (Friedman, 1991), have also been implemented for classification. Past research has indicated that such techniques may improve classification results in clinical settings, although results are not conclusive (Tu and Guerriere, 1993; Ripley, 1993; Ripley, 1994; Yarnold, 1995; Tu, 1996; and Duh *et al.*, 1998b). Analysis of a dichotomous outcome has been the most common application of neural nets in the statistical literature. Other applications, such as survival analysis (Faraggi and Simon, 1995), have been addressed elsewhere but are not considered in this study. Although appropriate for this study, methods for analysis of rates (with techniques comparable to Poisson regression) have not been developed for neural networks.

The field of occupational injury provides an excellent setting to examine the utility of neural networks for analyzing complex non-linear associations. Numerous measurements related to job activities and work exposures are often considered in assessing an individual's risk for injury (Hagberg *et al.*, 1997). Individual, physical workload, and organizational indices, as well as other occupational variables, have been linked to occupational injury (Burdorf *et al.*, 1997; Punnett, 1991). However, the nature of their association with injury is often difficult to define (Burdorf, 1992; Burdorf *et al.*, 1997; Hagberg *et al.*, 1997).

The goal of this study is to investigate neural networks specifically for identification of populations at high risk for occupational back injury. Due to the computational demands of the iterative procedures required for neural network analysis, the simulations are limited in terms of varying the simulation parameters, such as sample size, number of variables, and the network structure. Variations in these parameters could lead to different results and conclusions. Despite this restriction, this study makes a unique contribution to the neural net literature, as neural nets had not been previously applied to risk assessment in the field of occupational injury. In addition, most of the past publications, which utilized neural nets for prediction, were also limited in terms of varying network parameters (Lette *et al.*, 1994; Ripley, 1994; Loannidis *et al;* 1998) and utilized only a single data set to make conclusions (Tu and Guerriere, 1993; Koutsoukos *et al;* 1994; Lette, *et al.*, 1994; Lippman and Shahian, 1997; Duh *et al.*, 1998b; Loannidis *et al.*, 1998).

Distributions and associations were selected based on the appropriate literature to resemble (as closely as possible) true associations between injury and selected risk factors. As described in the methods, the specified covariates were chosen since their affect on the probability of back injury has been established but the exact nature of the (probably non-linear) associations is unclear, thus motivating the implementation of neural networks. For this study, simulations only address the question of whether neural nets can better identify populations at high risk for back injury when associations are not linear in a logit scale. These associations were limited to relationships considered realistic for ergonomic assessment measures and back injury, and therefore may not be the best examples of non-linearity appropriate for neural network analysis. Negative results should therefore not necessarily discourage use of neural models, but rather lead to further investigation of more optimal applications.

NEURAL NETWORK STRUCTURE

The basic unit of a neural network model is the semi-linear unit (Figure 1), which is similar to the logistic regression model. Define **s** as the network's inputs, which is simply the data for purposes of this study. The output, or response, of the semi-linear unit is determined by calculating the logistic function (Equation 1) of **x**, which is the vector product of the inputs and corresponding weights, **w**, so that **x** = **w** · **s** (Levine, 1991).

$$f(\mathbf{x}) = 1/\left(1 + \exp(-\mathbf{x})\right) \tag{1}$$

One distinction between the logistic regression model and the semi-linear unit is that the logistic function, in a semi-linear unit, is used to approximate the threshold function (*i.e.*, whether the neuron fires a signal). In logistic regression, the outcome is predicted as a linear function of the data in the logit scale (Hosmer and Lemeshow, 1989). Although the logit function is most commonly used, other functions, such as the probit function may be utilized. Ramifications of using a different function are not clear. The predicted prob-

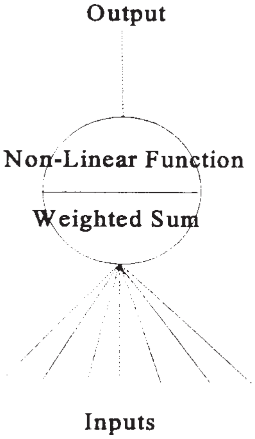

Figure 1. The semi-linear unit. (From Landsittel, D.P., Gardner, L.I., and Arena, V.C.)

ability from a neural net is not necessarily a linear function of the predictors and/or interactions (in any scale).

Other differences between the neural network model and logistic regression relate to the organization of semi-linear units into layers of the network. The neural network model is formed by connecting layers of semi-linear units, so that the outputs of units in one layer are used as inputs to the next layer (Levine, 1991). The initial layer of the network is the raw data and the layers between the first and last layers of the network are called hidden layers (since their outputs are hidden to the user). The final layer determines the response of the network, which in our case is the predicted probability of injury. In this study we utilize a network with one hidden layer and ten hidden units (Figure 2).

The purpose of utilizing hidden units is to transform the data into linearly separable groups (Levine, 1991). The output of the network can then be determined by classifying the data based on the weighted sum of the outputs of the hidden units. Interactions between variables and associations with the outcome are therefore implicitly determined by the network. It is unclear which types of associations in clinical or occupational settings are best classified by neural networks. The major trade-off between the two methods is that neural networks offer the flexibility of fitting non-linear associations without specifying the exact nature of the relationship, while logistic regression models offer the ability to specify exactly how the predictor variables interact with each other and the outcome.

Depending on the number of hidden units, the neural network model may include a much larger number of parameters than the logistic model. The total number of parameters in the network can be calculated as $H \cdot (K + 1) + H + 1$, where H denotes the number of hidden units (with 1 hidden layer and 1 output) and K denotes the number of variables in the model. For each hidden unit, a different weight (coefficient) is fitted for each variable and the intercept. For each output unit, a weight is also calculated for each unit in the hidden layer and the intercept in the hidden layer. For instance, a neural net

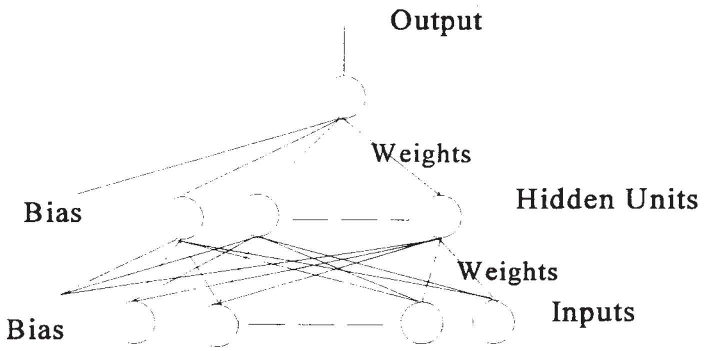

Figure 2. A network for anayzing binary outcomes. (From Landsittel, D.P., Gardner, L.I., and Arena, V.C.)

with 4 inputs and 10 hidden units will have 61 parameters to fit. The fully specified logistic model with 4 variables will have only 15 parameters to fit. These factors may lead to a greater possibility of overfitting random associations with neural networks. Theoretical considerations concerning the balance between maximizing accuracy and minimizing bias are published elsewhere (Geman, Bienstock, and Doursat, 1992).

As with standard regression methods, model complexity becomes especially problematic when the number of covariates is relatively large compared to the number of observations in the data set. Although no specific guidelines exist for calculating the required sample size of the training data set (in relation to model complexity), an adequate sample size for a standard regression method would not necessarily provide sufficient numbers for neural network analysis due to the greater number of parameters in the network model. Research in selecting of an adequate training set is not well developed. Other past publications discuss general considerations such as balancing precision and bias, and satisfying asymptotic properties (Geman *et al.*, 1992; White, 1989).

Optimal network weights can be determined through iterative numerical methods, such as back-propagation (Rumelhart *et al.*, 1995). Using random initial weights, the deviance (or error) of the model is calculated and the weights are updated based on a gradient descent learning rule. The process is continued until the deviance of the model is minimized. This procedure is often referred to as training the network. Since such routines may converge to a local minimum, we implemented two techniques, weight decay and committees of networks, to modify training (Ripley, 1995). Weight decay (Ripley, 1993) adds a penalty term to the deviance to improve convergence. With committees of networks (Rumelhart *et al.*, 1995), the predicted output is determined using five networks with different initial weights. The final output is calculated as the mean output of the five individual networks.

METHODS

For this analysis, we simulated data sets based on known associations between the outcome (back injury present or absent) and the predictor variables. The simulated data was generated to represent injury over a given period of time. Equal follow-up was therefore assumed. Simulated data sets were utilized to control for random associations which might occur in any single data sets. Repeated simulations allow us to better describe the true variability of classification results over many iterations. Simulating data also guarantees that the test data is truly independent of the training data. Any improvement in classification accuracy can then be attributed to specific known associations/data structures specified by the simulation conditions. Since neural nets are not restricted to a linear model in any scale, and implicitly fit interactions through the use of hidden units, overfitting is more likely than with logistic regression. In a simulation study we can assess classification accuracy under different known associations. In this simulation we specified both non-linear and completely random associations.

Based on the literature related to back injury risk factors, four different predictor variables (experience in years, body mass index in kg/m^2, percentage time spent lifting, and percentage of time in non-neutral trunk postures) were selected to be used in these simulations. The number of variables (four) in this study is limited by computational demands. These particular covariates were chosen since their effect on the probability of back injury has been established (see methods) but the exact nature of the (probably non-linear) associations is unclear, thus motivating the implementation of neural networks.

All variables were randomly generated from a multivariate normal distribution. Body mass index was generated independently ($\mu = 27$, $\sigma = 4$) from the other predictors. Experience ($\mu = 3$, $\sigma = 1$) was negatively correlated ($\rho = -0.7$) with both the percentage of time spent lifting and the percentage of time spent bending/twisting, implying that more experienced workers spend less time in material handling tasks. The percentage of time spent lifting ($\mu = 30$, $\sigma = 8$), and the percentage of time spent bending/twisting ($\mu = 30$, $\sigma = 8$) were generated with a positive correlation ($\rho = 0.7$). The distributions and parameters used to simulate these data were specified based on empirical frequency distributions from data sets currently being collected and analyzed. Simulated data was truncated at zero in the very rare instances were negative values were generated.

In the first set of simulations the association between the predictor variables and injury was completely random. Injury was randomly generated as a Bernoulli variable with probability of injury equal to 0.2, regardless of any covariate values. In these simulations we expected the classification accuracy to be no better than chance. Workers with predicted probabilities greater than 0.2 were classified as high risk for injury. One thousand (training) data sets, each with a sample size of 100, were randomly generated and used to fit the logistic and neural network models. The percentage of individuals correctly classified, and the percentage of false positives and false negatives were reported using both methods. Confidence intervals were determined by the 5th and 95th percentiles of the results from the 1000 simulations. To assess generalization of these models, an additional 1000 (test) data sets were generated using the same distributions and association. Classification results for the test data were reported using the models fit with the training data.

In the second set of simulations, a non-linear association was specified between injury and each of the risk factors. The relationships used to generate these associations were motivated by findings in the relevant epidemiologic and ergonomic literature. The underlying probability of injury for each worker was calculated using the following assumptions. Injury status was randomly generated from a Bernoulli distribution with the specified probability of injury.

1. Less experienced workers experience higher back injury rates (Kelsey and Golden, 1988; Kraus *et al.*, 1996).

2. Workers with less than, or greater than optimal body mass index experience higher back injury rates (Kelsey and Golden, 1988).

3. Workers who lift frequently experience higher back injury rates (Burdorf, 1992; Kraus, *et al.*, 1996).

4. Workers who bend or twist frequently experience higher back injury rates (Burdorf, 1992; Kelsey and Golden, 1988; Punnett *et al.*, 1991).

5. Workers who lift and bend or twist frequently experience an interactive effect.

We specified a baseline risk of 0.05 to generate the probability of back injury for each individual worker. To incorporate the previously mentioned assumptions, each worker's risk for back injury was increased by some increment for the level of each risk factor present. For instance, the risk of back injury was additively increased by 0.10 for workers with less than 1 year of previous experience, and by 0.05 for workers with less than 2 years of experience (based on assumption 1 above). Increased risks for different levels of each variable are listed in Table 1. For individuals with a BMI greater than 30, or less than 20, the additional risk of back injury increases linearly with an increase in BMI. Similar associations are specified with percent of time spent lifting and percent of time spent bending or twisting. An interactive effect was simulated for individuals who lift, and bend or twist frequently (greater than 30% of the time). The magnitude of this effect, which is described in Table 2, differs depending on the level of the worker's body mass index and experience. The increased risk is highest for workers with higher than optimal body mass index and less than one year of experience.

Based on each individual's covariates, their probability for back injury was calculated using the associations listed in Table 1 and Table 2. Each worker's injury status was then randomly generated from a Bernoulli distribution with the appropriate probability of injury. Workers with a predicted probability of injury greater than 0.2 were again classified as high risk. One thousand (training) data sets, each with a sample size of 100, were randomly generated and used to fit the logistic and neural network models. Classification tables were calculated with both neural networks and logistic regression for each simulation. To assess generalization of these models, an additional 1000 (test) data sets were generated using the same distributions and association. Classification results for the test data were reported using the models fit with the training data.

Two different logistic regression models were used in each simulation to identify high-risk populations. We specified both the main effects model and the full model with all possible interactions. Since the inclusion / deletion of

Table 1. Simulated increases in the probability of back injury by risk factor.

Variable	Category	Increased risk
Previous experience (in years)	<1	0.10
	≥1, <2	0.05
Body mass index (BMI)	<20	(20-BMI)/100
	≥20, <30	0
	≥30	(BMI-30)/100
% Lifting (%L)	>30	(%L-30)/100
% Bending/twisting (%B)	>30	(%B-30)/100

Table 2. Simulated Interaction between risk factors.

Body mass index (BMI)	Experience (in years)	Increased risk
<20	<1	0.08
	≥1, <2	0.06
	≥2	0.04
≥20, <30	<1	0.06
	≥1, <2	0.04
	≥2	0.02
≥30	<1	0.10
	≥1, <2	0.08
	≥2	0.06

interaction terms cannot be controlled in neural networks, the full logistic model provides the closest possible comparison. The main effects model is also fit for illustrative purposes. Results of other logistic models are not relevant to this comparison. Both the logistic and neural net models were fit using continuous variables.

Since the simulated data structure is completely known, a logistic model could be fit to model the exact associations which are used to generate the data. Such an analysis would undoubtedly produce more accurate results with logistic regression. The specific objective of these simulations, however, was to assess the ability of neural networks to identify high risk populations in the case where the underlying data structure is unknown. Therefore, for the purposes of this study, the covariate values were left as continuous, and more complex parameters were left out of the logistic model.

The data sets generated here can only be considered linear and multiplicative if the cut points for determining probability of injury are known a priori (*i.e.*, if the underlying structure of the data is known). For the purposes of this study, the analysis conducted does not assume any such knowledge and therefore treats the associations as unknown and essentially non-linear. The variables were therefore specified as strictly continuous in each model. Truly non-linear associations were not attempted. The authors limited the simulations to associations which could be easily justified as realistic in this setting.

RESULTS

In the first set of simulations, we analyzed data generated from a completely random association. The percentage correctly classified (%CC), percentage of false positives (%FP), and percentage of false negatives (%FN) were reported (Table 3). Neural networks correctly classified 80% of the injured workers in the training set as high risk. The main effects and fully specified logistic models correctly classified 60 and 70%, respectively, of the observations. Neural networks correctly classified injury status in 55% of the workers in the test data, as compared to 80% of the workers in the training data. The main effects and fully specified logistic models correctly classified 52 and 56%, respectively, of the observations in the test data. Results concerning the test set of random associations indicate that, for 1000 simulations of sample size 100, considerable variability in classification results exists. Given no true association in an independently generated data set, one would expect 50% of the injured to be classified as high risk.

In the next set of simulations, we analyzed data generated from the previously described non-linear associations. The percentage correctly classified (%CC), percentage of false positives (%FP), and percentage of false negatives (%FN) were reported (Table 4). Neural networks correctly classified an average of 85% of the workers in the training data. The main effects and fully specified logistic models correctly classified 74 and 79%, respectively, of the observations. The percentage of workers correctly classified with neural networks was again substantially lower using the test data (67%), as compared to results using the training data (85%). The main effects and fully specified logistic models correctly classified 71 and 69%, respectively, of the observations in the test data.

DISCUSSION

Neural networks have been implemented in past studies to identify high-risk populations in the health care setting (Duh *et al.*, 1998b; Koutsoukos *et al.*, 1994; Tu and Guerriere, 1993). This study represents the first application of neural networks to occupational injury epidemiology. Results from our simulations show that neural networks do not provide any benefit over standard statistical methods in identifying populations at high risk for back injury for these particular simulation conditions, *i.e.*, non-linear associations with a lim-

Table 3. Classification of completely random associations.

		%CC		%FP		%FN	
Model	**Data**	**Mean**	**95% CI**	**Mean**	**95% CI**	**Mean**	**95% CI**
Main effects-logistic	Training	0.60	0.36,0.80	0.73	0.65,0.82	0.14	0.09,0.21
	Test	0.52	0.28,0.74	0.80	0.68,0.91	0.20	0.10,0.30
Full model-logistic	Training	0.70	0.53,0.83	0.64	0.53,0.73	0.10	0.05,0.15
	Test	0.56	0.40,0.70	0.80	0.69,0.91	0.20	0.12,0.29
Neural network	Training	0.80	0.64,0.93	0.48	0.29,0.61	0.02	0.00,0.07
	Test	0.55	0.38,0.70	0.80	0.69,0.90	0.29	0.12,0.20

Table 4. Classification of non-linear associations.

		%CC		%FP		%FN	
Model	Data	Mean	95% CI	Mean	95% CI	Mean	95% CI
Main effects-logistic	Training	0.74	0.62,0.86	0.64	0.52,0.75	0.09	0.04,0.13
	Test	0.71	0.57,0.82	0.68	0.52,0.84	0.11	0.04,0.18
Full model-logistic	Training	0.79	0.68,0.90	0.55	0.42,0.67	0.06	0.03,0.10
	Test	0.69	0.57,0.79	0.73	0.58,0.89	0.13	0.06,0.21
Neural network	Training	0.85	0.74,0.95	0.45	0.27,0.58	0.01	0.00,0.04
	Test	0.67	0.54,0.79	0.73	0.58,0.87	0.12	0.04,0.19

ited sample size and a limited number of variables. These findings are consistent with recent epidemiologic studies (Duh *et al.*, 1998b; Ripley, 1994), which indicated that neural networks may provide only equivalent prediction results to logistic models and other regression techniques for common biostatistical settings.

In simulations of a completely random association, neural networks, despite the lack of any true relationship between predictors and injury status, correctly identified injured workers as high-risk 80% of the time. Utilizing even the fully specified logistic model led to results much closer to those expected under a completely random association. For the non-linear association, when using the actual training data to obtain classification results, the percentage correctly classified with neural networks appears slightly higher (although the CIs overlap) than the percentage obtained with logistic regression (Table 4 — 85% vs. 79% and 74%). When using the test data, however, the percentage correctly classified is slightly greater with logistic regression (71 and 69% vs. 67%). The network model likely overfits the true association since neural networks do not allow the user to specify the nature of the association or exclude interaction terms from the model.

Results underscore the importance of using cross-validation methods to modify network training based on test set data and / or obtaining classification results on the independent test set. (Cross-validation methods were not implemented here due to the extremely computational nature of the simulations.) Even though the differences between results using the training set and results using the test set were much greater with neural nets, classification of the training data is still biased with logistic regression. Using the full model, the percent correctly classified was 10 to 15% greater with the training data. Even using the main effects model the percent correctly classified was greater with the training data (although not substantially different). Thus, although such issues are most relevant to neural nets, results serve to motivate the use of an independent test set for prediction with even standard statistical methods. The percentage of false positives and false negatives using the test data was very similar with each of the three models.

Future studies should address the ability of neural networks to identify high-risk populations under different assumptions. In this study, we generated data sets from a particular multivariate normal distribution and a single set of assumptions regarding the association between the predictor variables and injury status. The sample size and number of variables were held constant. Additional simulations should examine variations in these parameters. Very few studies have quantitatively addressed the effect of variations in network parameters or attempted to evaluated the utility of neural nets for different types of data sets (Duh *et al.*, 1998a). Analysis done by Ripley (1995), for instance, indicates little difference between neural nets fit using 3, 5, or 8 hidden units, although weight decay had a significant effect on model fit. Results were only demonstrated for a single data set with 100 observations. Although other studies have done some analysis to determine the relative importance of such factors (Duh *et al.*, 1998a; Duh *et al.*, 1998b; Lippmann and

Shahian, 1997; Ripley, 1994; Geman *et al.*, 1992), these studies do not provide clear guidelines regarding the inter-relationship between model structure, sample size, and other factors in risk assessment. Using the results of this study, researchers should run further analyses / simulations to investigate such issues.

Some generalizations can be made about possible / probable ramifications. A greater number of hidden units increases the network's ability to transform the data and achieve linear separability (Hertz *et al.*, 1991; Levine, 1991). Specifying networks with more hidden units or hidden layers (*i.e.*, more parameters to fit) would therefore likely lead to more accurate results in terms of identifying injured workers as high risk, although overfitting will likely worsen (Landsittel, 1997).

For the purposes of this analysis we generated injury status assuming equal follow-up. In most cases, where differential follow-up exists, the injury rates (rather than injury status) are analyzed. Currently, methods to accomplish this with neural networks are not available, although several publications have recently addressed implementing neural nets for survival analysis (Faraggi and Simon, 1995; Liestol, Anderson, and Anderson, 1994). Estimating levels of risk associated with particular categories of a given variable is also very difficult with neural networks. Because the weights associated with hidden units are not directly interpretable as coefficients for a particular variable, assessing the magnitude and direction of a particular variable's effect on the outcome is problematic.

Results of this study do not completely answer the research question of when to use neural networks for occupational injury risk assessment. The results do, however, indicate that departures from linearity in a logit scale do not provide sufficient motivation for implementing neural network analysis. Researchers should first decide whether the underlying structure of the data can be reasonably described through conventional methods. If so, standard regression methods will likely provide adequate results. In situations where this cannot be accomplished, neural networks may still prove useful since the user only specifies which variables are included in the model. Further investigations of specific conditions where neural nets are most useful are needed.

REFERENCES

Anderson, T.W. 1984. *An Introduction to Multivariate Statistical Analysis.* New York, NY: John Wiley & Sons.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees.* Belmont, CA: Wadsworth International.

Burdorf, A. 1992. Exposure assessment of risk factors for disorders of the back in occupational epidemiology. *Scandinavian Journal of Work Environment and Health* **18,** 1–9.

Burdorf, A., Rossignol, M., Fathallah, F.A., Snook, S.H., Herrick, R.F. 1997. Challenges in assessing risk factors in epidemiologic studies on back disorders. *American Journal of Industrial Medicine* **32,** 142–152.

Courtney, T.K., Burdorf, A., Sorok, G.S., and Herrick, R.F. 1997. Methodological Challenges to the study of occupational injury - An international epidemiology workshop. *American Journal of Industrial Medicine* **32,** 103–106.

Duh, M., Walker, A., and Ayanian, J. Z. 1998a. Epidemiologic interpretation of artificial neural networks. *American Journal of Epidemiology* **147,** 1112–1122.

Duh, M., Walker, A.M., Pagano, M., and Kronlund, K. 1998b. Prediction and cross-validation of neural networks versus logistic regression: Using hepatic disorders as an example. *American Journal of Epidemiology* **147(4),** 407–413.

Faraggi, D. and Simon, R. 1995. A neural network model for survival data. *Statistics in Medicine* **14,** 73–82.

Friedman, J.H. 1987. Exploratory projection pursuit. *Journal of the American Statistical Association* **82(397),** 249–266.

Friedman, J.H. 1991. Multivariate adaptive regression splines. *Annals of Statistics* **19,** 1–141.

Geman, S., Bienenstock, E., and Doursat, R. 1992. Neural networks and the bias/variance dilemma. *Neural Computation* **4,** 1–58.

Hagberg, M., Christiani, D., Courtney, T.K., Halperin, W., Leamon, T.B., and Smith, T.J. 1997. Conceptual and definitional issues in occupational injury epidemiology. *American Journal of Industrial Medicine* **32,** 106–115.

Hertz, J., Krogh, A., and Palmer, R.G. 1991. *Introduction to the Theory of Neural Computations*. Reading, MA: Addison-Wesley Publishing Company.

Hosmer, D.W. and Lemeshow, S. 1989. *Applied Logistic Regression*. New York, NY: John Wiley & Sons.

Jones, M.C. and Sibson, R. 1987. What is projection pursuit? *Journal of the Royal Statistical Society A* **150,** 1-36.

Kelsey, J.L. and Golden, A.L. 1988. Occupational and workplace factors associated with low back pain. *Occupational Medicine: State of the Art Reviews* **3(1),** 7–16.

Koutsoukos, A.D., Rubinstein, L.V., Faraggi, D., Simon, R.M., Kalyandrug, S., Weinstein, J.N., Kohn, K.W., and Paull, K.D. 1994. Discrimination techniques applied to the NCI in virto anti-tuour drug screen: predicting biochemical mechanism of action. *Statistics in Medicine* **13,** 719–730.

Kraus, J.F., Brown, K.A., McArthur, D.L., Peek-Asa, C., Samaniego, L., Kraus, C, and Zhou, L. 1996. Reduction of acute low back injuries by use of back supports. *International Journal of Occupational Environment and Health* **2,** 264–273.

Landsittel, D. 1997. *A Simulation Study of Statistical Modeling with Neural Networks*. Ph.D. dissertation. Department of Biostatistics, University of Pittsburgh.

Levine, D.S. 1991. *Introduction to Neural and Cognitive Modeling*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Liestol, K., Anderson, P.K., and Anderson, U. 1994. Survival analysis and neural nets. *Statistics in Medicine* **13,** 1189–1200.

Lippmann, R. P. and Shahian, D. M. 1997. Coronary artery bypass risk prediction using neural networks. *Annals of Thoracic Surgery* **63,** 1635–1643.

Loadnnidis, J. P., McQueen, P. G., Goedert, J. J., and Kaslow, R. A. 1998. Use of neural networks to model complex immunogenetic associations of disease: human leukocyte antigen impact on the progression of human immunodeficiency virus infection. *American Journal of Epidemiology* **147,** 464–471.

Myers, R.H. 1990. *Classical and Modern Regression with Applications*. Boston, MA: PWS-Kent Publishing Company.

Punnett, L., Fine, L.J., Keyserling, W.M., Herrin, G.D., and Chaffin, D.B. 1991. Back disorders and nonneutral trunk postures of automobile assembly workers. *Scandinavian Journal of Work Environment and Health* **17,** 337–346.

Ripley, B.D. 1993. Statistical aspects of neural networks. In: *Networks and Chaos — Statistical and Probabilistic Aspects*, pp. 40–123. (Barndorff-Nielsen, O.E., Jensen, J.L., and Kendall, W.S., Eds.) New York, NY: Chapman & Hall.

Ripley, B.D. 1994. Neural networks and related methods for classification. *Journal of the Royal Statistical Society B* **56(3),** 409–456.

Ripley, B. D. 1995. Statistical ideas for selecting network architectures', In: Kappen, B. and Gielen, S. (Eds), *Neural Networks: Artificial Intelligence and Industrial Applications*, Springer, London, 183–190.

Rumelhart, D.E., Durbin, R., Golden, R., and Chauvin, Y. 1995. Backpropagation: the basic theory. In: *Backpropagation: Theory, Architectures, and Applications*, pp. 1–34. (Chauvin, Y., and Rumelhart, D.E., Eds.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Schley, C., Chauvin, Y., and Henkle, V. 1995. Automated aircraft flare and touchdown control using neural networks. In: *Backpropagation: Theory, Architectures, and Applications*, pp. 1–34. (Chauvin, Y., and Rumelhart, D.E., Eds.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Stern, H.S. 1996. Neural networks in applied statistics. *Technometrics* **38(3)**, 205–218.

Tu, J.V. and Guerriere, M.R. 1993. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. *Computers and Biomedical Research* **26,** 220–229.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. 1995. Phoneme recognition using time-delay neural networks. In: *Backpropagation: Theory, Architectures, and Applications*, pp. 1–34. (Chauvin, Y., and Rumelhart, D.E., Eds.) Hillsdale, NJ: Lawrence Erlbaum Associates.

White, H. 1989. Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association* **84,** 1003–1013.

Yarnold, P.R., Soltysik, R.C., McCormick, W.C., Burns, R., Lin, E.H, Bush, T., and Martin, G.J. 1995. Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine* **10(11),** 601–606.