ELSEVIER

Accident Analysis and Prevention 38 (2006) 434–444

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks

Dursun Delen*, Ramesh Sharda, Max Bessonov

Department of Management Science and Information Systems, Oklahoma State University, Stillwater, OK 74106, United States

Received 25 April 2005; received in revised form 26 May 2005; accepted 15 June 2005

Abstract

Understanding the circumstances under which drivers and passengers are more likely to be killed or more severely injured in an automobile accident can help improve the overall driving safety situation. Factors that affect the risk of increased injury of occupants in the event of an automotive accident include demographic or behavioral characteristics of the person, environmental factors and roadway conditions at the time of the accident occurrence, technical characteristics of the vehicle itself, among others. This study uses a series of artificial neural networks to model the potentially non-linear relationships between the injury severity levels and crash-related factors. It then conducts sensitivity analysis on the trained neural network models to identify the prioritized importance of crash-related factors as they apply to different injury severity levels. In the process, the problem of five-class prediction is decomposed into a set of binary prediction models (using a nationally representative sample of 30 358 police-recorded crash reports) in order to obtain the granularity of information needed to identify the "true" cause and effect relationships between the crash-related factors and different levels of injury severity. The results, mostly validated by the findings of previous studies, provide insight into the changing importance of crash factors with the changing injury severity levels.

Keywords: Injury severity; Classification; Artificial neural networks; Sensitivity analysis; Problem decomposition

1. Introduction

Over 6 million traffic accidents claim more than 40 000 lives each year in the United States according, to the National Highway Traffic Safety Administration (GES, 2005). Causes of accidents and related injury severity are of special concern to researchers in traffic safety, since such research would be aimed not only at prevention of accidents but also at reduction of their severity. One way to accomplish the latter is to identify the most probable factors that affect injury severity. Understanding the circumstances under which drivers and passengers are more likely to be killed or more severely injured in an automobile accident can help improve the overall driving safety situation. Factors that affect the risk of increased injury of occupants in the event of an automotive accident include demographic or behavioral characteristics of the person (age, gender, seatbelt usage, or use of drugs or alcohol while driving), environmental factors and roadway conditions at the time of the accident occurrence (surface, weather or light conditions, the direction of impact, vehicle role, or occurrence of a rollover), as well as technical characteristics of the vehicle itself (vehicle age and body type).

The primary interest of this study is to identify which of these factors become important in influencing the probability of increased injury severity during a crash. Accidents examined herein include a geographically representative sample of multi-vehicle collision accidents, single vehicle fixed-object collisions, and single vehicle non-collision (rollover) crashes. Many of the previous studies in this domain have used regression type generalized linear models where the functional relationships between the injury severity and the crash-related factors are assumed to be linear. As noted by Mussone et al. (1999), these linear models suffer from problems related to the use of variables with non-homogenous distribution and are known to be prone to statistical problems when the correlation among the independent variables is greater than acceptable levels, resulting in unreliable models with greater error than desirable. Since artificial neural networks are capable of capturing highly non-linear relationships between the predictor variables (crash factors) and the target variable (severity level of the injuries), in this study, a series of artificial neural networks models were

* Corresponding author. Tel.: +1 918 594 8283; fax: +1 918 594 8281.
E-mail address: Delen@okstate.edu (D. Delen).

0001-4575/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2005.06.024

used to estimate the effects of significant crash factors on the level of injury severity sustained by the driver. Though artificial neural networks have been investigated by other researchers in this domain (Abdelwahab and Abdel-Aty, 2001; Mussone et al., 1999), we propose and employ some methodological innovations to develop the sensitivity analysis from the neural network models.

This paper is organized as follows. The next section summarizes previous research in accident analysis. Section 3 describes our data compilation and processing approach, as well as definition of all the variables. Section 4 presents the details of the experiments to determine the effect of the various independent variables. Section 5 includes the results of these experiments along with the discussion of the findings, and the last section summarizes and concludes the paper.

2. Prior research

There is an extensive literature devoted to analysis of traffic accidents and crash modeling. The majority of this literature references the studies that deal primarily with analysis of crash involvement and prediction of crash totals. Since this research deals with analysis of injury severity and its deriving factors, this review attempts only to present the most relevant and rigorous work in this area.

A number of studies have attempted to develop injury severity models, though most concentrate on the traffic accident records limited to a small geographic area, a particular accident type, or a certain road condition situation. The reason is to make the domain as narrow as possible so that a somewhat homogenous dataset can be used to derive accurate prediction models. What follows is a review of a number of recent studies that develop models to identify the factors most important in increasing or reducing the levels of injury severity experienced by drivers and/or passengers during traffic accidents.

Logistic regression (a type of regression where the dependent is a categorical as opposed to a numerical variable) has been the most popular technique in developing injury severity prediction models. Lui and McGee (1988) used logistic regression to analyze the probability of fatal outcomes of accidents given that the crash has occurred. The data were obtained from the Fatal Accident Reporting System, which contained accidents that included at least one fatality. Probability of a fatal outcome was modeled as a target variable, which is dependant on driver's age and gender, impact points, car deformation, use of restraint system, and vehicle weight. Their findings reveal that a heavier car weight can greatly reduce the driver's risk of dying in a two-car crash, credited to the intuition that larger cars' frames can better absorb energy from an impact, or the fact that the small and lighter cars tend to roll over more easily. These findings were also confirmed by Wood and Simms' (2002) study where they measure the impact of car size on injury risk. Lui and McGee (1988) also analytically confirmed that the weakest side of a car from a driver's standpoint is the left side, and a belted driver subject to a left front side impact is estimated to have a higher risk of dying then an unbelted driver.

In another recent study a multivariate logistic regression (Bedard and Lu, 2002) revealed that the odds ratio of a fatal outcome of a crash increases with age, reaching 4.98 for drivers aged 80+ compared to the drivers aged 40–49 years. Female gender and blood alcohol content greater than 0.30 were also to be found associated with higher fatality odds. Also, the driver side impacts doubled the odds of fatality compared to frontal impacts. This study was limited to single vehicle crashes with fixed objects. In yet another study a logistic regression approach was used to examine a contribution of individual variables to the injury severity (Al-Ghamdi, 2002). The study was limited to 560 accidents obtained from the police records in Riyadh, Saudi Arabia. The dependent variable was modeled as a dichotomous variable that could only take values of fatal or non-fatal crash outcomes. According to the logistic regression results, out of nine independent variables used in this study, only two were found to be statistically significant with respect to the injury severity: location and cause of accident. For example, the odds of being in a fatal accident at a non-intersection location are 2.64 higher than those at an intersection, and the odds of severe injury increases on accidents caused by over-speeding and entering the wrong way traffic.

Farmer et al. (1997) used a binomial regression model to investigate the impact of vehicle and crash characteristics on injury severity in two-vehicle side-impact crashes. They found that rollover or ejection from the vehicle increases the likelihood of a serious injury or death (i.e., injuries with an Abbreviated Injury Scale of at least 3) and that light-duty trucks (which they defined as pickups and utility vehicles – not vans – under 10 000 pounds of gross-vehicle weight [GVW]) were fourteen times more likely to roll than cars, when struck on the side. While gender was not a statistically significant factor in their results, the oldest drivers (aged 65 and over) were estimated to be more at risk for serious injury.

Recently, ordered probit models have become popular in analysis of multi-class injury severity data sets. For instance, an ordered probit model was employed at the University of Texas (Kockelman and Kweon, 2002) in order to examine the risk of different injury levels sustained under a variety of crash types. The results suggest that pickups and SUVs are less safe than passenger cars under single-vehicle accident conditions. However, in two-vehicle accidents, they were found to be safer for their drivers and more dangerous for the occupants of their collision partners. In a similar study, using 1994 and 1995 crash data from the state of Florida, Abdel-aty et al. (1998) used ordered probit technique to examine relationships between driver age and crash characteristics. The three injury severities in their study were no injury, injury and fatality, and their results suggest that injury severity is positively associated with age. They also concluded that middle-age drivers are more likely to be involved in some crashes, but older drivers are more likely to be involved in fatal crashes.

O'Donnell and Connor (1996) compared ordered logit and ordered probit models in assessing the probabilities of four levels of injury severity as a function of driver attributes. In logit and probit regression models, while predicting the categorical response variable, a membership probability value (in the form of a continuous function that naturally stays within the 0–1 bounds) is generated. The research results suggested that injury severity rises with speed, vehicle age, occupant age (squared), female gender, blood alcohol levels over 0.08%, non-use of a seatbelt, manner of collision (e.g., head-on crashes), and travel in a light-duty truck. According to their comparison, seating position of crash victims was the most relevant (e.g., the left-rear seat of the vehicle was found to be the most dangerous) and gender the least relevant factor. Many of their results are verified in the models presented by us; the key distinction is that in this paper collision partners and crash-type are examined and emphasized. Also, our results are based on a much larger data set.

Use of artificial neural networks as the modeling approach in analysis of crash-related injury severity has been relatively scarce. The only related study that was found in the literature employed artificial neural networks to model the relationship between driver injury severity and crash factors related to driver, vehicle, roadway, and environment characteristics (Abdelwahab and Abdel-Aty, 2001). Their study focused on classifying an accident into one of three injury severity levels using the readily available crash factors. They limit their domain of study to two-vehicle accidents that occurred at signalized intersections. The predictive performance of multi-layer perceptron (MLP) neural network was compared to the performance of the ordered logit model. Their results revealed that MLP was a better classifier (correctly classifying 65.6 and 60.4% of cases for the training and testing phases, respectively) than ordered logit model (correctly classifying 58.9 and 57.1% of cases for the training and testing phases, respectively) for the specific data being analyzed.

We differentiate our work from those of the ones referenced in this section by two dimensions. First, we used artificial neural networks as the modeling medium to capture the potentially non-linear relationships between the injury severity levels and crash-related factors. We then developed and used sensitivity analysis on those trained neural network models to identify the prioritized importance of crash-related factors as they apply to the injury severity levels. Second, as opposed to aggregating the five levels of injury severity to two or three aggregate levels for a single prediction model (as has been the case in most of the previous studies), we used a partitioning method where we have converted the five-class prediction model into eight binary classification models. This gave us the granularity that we seeked to obtain in identifying the "true" cause and effect relationships between the crash-related factors and different levels of injury severity.

3. Data

This research investigates the injury severity experienced by drivers in automobile crashes without limiting the study to any specific geographic area of the United States. Therefore, data were acquired from the National Automotive Sampling System General Estimates System, which covers approximately 0.85% of all US crashes reported by the police (GES, 2005). This dataset is compiled as a nationally representative sample of all police crash reports. The reports contain data on all property damage crashes as well as injury-causing and fatal crashes. Although GES suggests that many motor vehicle accidents are not reported to the police, majority of these unreported cases involve only negligible property damage and no significant injury to people involved. Thus, restricting data collection to police-reported crashes, GES concentrates on crashes with greatest influence on roadway safety. This study used 30 358 police-reported accident records (a significantly larger data set compared to other studies reported in the previous literature) that describe motor vehicle crashes from 1995 to 2000. This 30 358 record dataset does not include the records removed during the pre-processing.

The GES dataset was obtained in three separate files: accident, vehicle, and person record sets on a year-by-year basis. The *accident files* contained information about the environment, road conditions, and the circumstances of the crash. The *vehicle files* described various characteristics of the vehicle(s) involved. The *person files* provided information on all the people involved in the accident. Each accident described in the database involved one or more vehicles, and each vehicle contained one or more occupants. For the purposes of this study, the three data files for each of the 5 years were combined to create a single dataset of person-level records containing all relevant data about the drivers involved in those crashes. All of the records used in study reported an injury level sustained by the driver as a result of an accident. These injury levels were encoded as no injury, possible injury, minor non-incapacitating injury, incapacitating injury, and fatality. Compared to abbreviated injury scale (AIS), a commonly used general purpose injury scale, the five-scale injury levels used in this study (INJ_SEV) are more aggregated where the category 1 in AIS roughly corresponds to category 2 of INJ_SEV, category 2 and 3 in AIS roughly corresponds to category 3 of INJ_SEV, category 4 and 5 in AIS corresponds to category 4 of INJ_SEV, category 6 in AIS corresponds to category 5 of INJ_SEV, and there is corresponding category in AIS for category 1 (no injury) in INJ_SEV. Accidents that did not result in any damage to the vehicle were removed from the study by using GES variable VEH_SEV, which reports the severity of the vehicle damage. The injuries resulting form the accidents without at least minor damage to the vehicle were very rare in occurrence. Therefore, this study examined crash records that exhibited minor, functional (moderate), or severe (vehicle rendered inoperable) damage to the vehicle. Since the study was concerned with identifying significant predictors that affect the injury severity levels of the drivers involved in crashes, it ignored all accidents involving pedestrians. The final dataset contained (1) single vehicle non-collision (rollover) crashes, (2) single vehicle fixed-object collisions, and (3) multi-vehicle collision accidents. This study also concentrated on accidents involving passenger cars, SUVs, vans, and light trucks and excluded records that involved motorcycle, bus, farm equipment, and heavy truck accidents.

Approximately 9.5% of all records in the GES dataset contained missing values for at least one variable. NHTSA imputes some of the missing variable values using a simple univariate method or a more sophisticated hot-deck method. No matter how sophisticated the imputation technique is, it is a fact that the imputed values are not real, and may introduce bias. It is advisable to remove such cases when (i) there is enough data left for analysis and (ii) the removal of these cases does not introduce bias into the dataset (i.e., affecting the distribution of variables). After careful consideration, this study excluded those records and only included records without any missing data. Preliminary tests, conducted as part of this study, showed that this exclusion did not significantly change the distributions of the variables and hence did not bias the experiment results.

Along with commonly used techniques such as Chi-squared test, stepwise logistic regression and decision tree induction, the existing research literature was consulted in selecting the independent variables for this study. While the GES dataset contains over 150 variables, most of these variables are only related to very few specific cases out of the overall collection. Some variables, such as weather conditions, failed to become significant predictors of injury severity in any of the preliminary experimental runs. Excluding weather conditions from the set of independent variables is also in agreement with the previous research (Edwards, 1998). Apparently, most drivers adjust their driving style in accordance with worsened weather conditions, thus countering the higher risk of accident occurrence and injury possibility. Out of all the variables provided by the GES database, 17 were selected as important in influencing the level of injury severity of driver involved in accidents. These variables provide information about the driver and vehicle involved in the crash, the environmental conditions surrounding the crash, the circumstances of the accident, and additional information about when the accident took place. All of the variables except for AGE (age of the driver) and VEH_AGE (age of the vehicle) are categorical. Table 1 describes all the variables selected for this study, provides the interpretation of values used for coding of categorical variables, and shows the number of examples for each categorical value.

The variables (both quantitative and categorical) are then normalized using the min-max normalization formula (see Eq. (1)) where the minimum of 0 and maximum of 1 are used to match with the upper and lower limits of the sigmoid activation function used in the ANN models. Non-binary categorical variables (having more than two possible values) are converted into 1-of-N binary representation prior to performing the normalization.

$$v' = \frac{v - \min_A}{\max_A - \min_A}(\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A \qquad (1)$$

where v' is the normalized value, v is the original value, $\max_A$ and $\min_A$ are the maximum and minimum values of the original variable A, new_max_A and new_min_A are the maximum and minimum values of the min-max normalization range (1 and 0, respectively).

4. Method and experiment

The purpose of this study is to identify and prioritize the accident, vehicle, and person related factors that contribute to the increased severity of an injury during traffic accidents. Since the dependent variable (injury severity) is presented as a discrete variable with five possible outcomes, the problem has become one of classification.

4.1. Neural networks

Neural networks are commonly known as biologically inspired, highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions. Formally defined, neural networks are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data (Haykin, 1998). We used a popular neural network architecture called multi-layer perceptron (MLP) with back-propagation gradient-descent supervised learning algorithm and sigmoid activation functions for the processing elements (PE) (a.k.a. perceptrons or neurons). An MLP is essentially the collection of non-linear neurons organized and connected to each other in a feedforward multi-layer structure using directed arrows as coefficients (commonly called as weights in neural network terminology). The MLP is known to be a robust function approximator for prediction and classification problems. It is arguably the most commonly used and well-studied NN architecture. Hornik et al. (1990) empirically showed that given the right size and the structure, an MLP is capable of learning arbitrarily complex non-linear functions to an arbitrary accuracy level. Fig. 1 shows the graphical representation of the original MLP used in this study. The subsequent neural network models (i.e. the eight binary classification models) employed the same structure with the exception of having the binary output variable.

We used a cross-validation based experimental design in determining the "optimal" number of hidden neurons (or we probably should say "sub-optimal" since it is not feasible to try every possible combination of options for a global optimum). Specifically, we started from a small number of neurons, measured the performance of the trained NN on holdout samples; gradually increased the number of neurons at a level where the performance of the trained NN on holdout samples is started to deteriorate due to overtraining, and took that as the indication of "optimal" number of neurons for the hidden layer of NN (18 for this study). In order to decrease the bias associated with the random splitting of the dataset into training and testing, we used a 10-fold cross-validation on each step of the experimental analysis explained herein.

The categories of different injury levels are not equally represented in the dataset. Table 2 shows the number of cases in each injury category. As can be seen, the *no injury* category contains 10 times the number of cases of *fatality* category. When this type of highly skewed data is introduced to a classification algorithm, the model would simply ignore the less represented categories in order to improve the overall accuracy. However, it is often more important to predict the less representative classes than having better overall accuracy. According to Wilson and Sharda (1994), better-balanced training data lead to better predictive accuracy in ANNs. With such skewed data, our ANN did not produce

Table 1
Injury severity levels represented in the dataset

Variable type	Variable name	Possible values and number of examples
Dependent variable	INJ_SEV	(0) No injury (10748) (1) Possible injury (7166) (2) Non-incapacitating injury (5941) (3) Incapacitating injury (5408) (4) Fatality (1095)
Person information	AGE	Age of the person in years
	SEX	(0) Female (10644) (1) Male (19714)
	ALC_DRU	(1) If the driver is under influence of drugs and/or alcohol (1027) (0) If not (29331)
Vehicle information	VEH_AGE	Vehicle age in years (derived from the model year)
	BODY_TYP	(0) Passenger cars (18254) (1) SUVs (1899) (2) Vans (2248) (3) Pickup/light trucks (7957)
	REST_SYS	(0) No restraint system used (1878) (1) Restraint system used (lap belt, shoulder belt, or both) (28480)
Environment information	INT_HWY	(1) If the accident happened on an interstate highway (2201) (0) If not (28157)
	LGHT_CON	(1) Daylight (including dusk and dawn) (21935) (0) No daylight (8423)
	SUR_COND	(0) Normal surface (21742) (1) Slippery surface (8616)
Accident info	STRIKING	Indicates vehicle role in single or multi-vehicle crashes. Derived from VEH_ROLE. Vehicles that were both striking and struck will have 1 as a value of both variables, and vehicles that did neither have 0 for both variables (0) No (14647) (1) Yes (15711)
	STRUCK	Indicates vehicle role in single or multi-vehicle crashes. Derived from VEH_ROLE. Vehicles that were both striking and struck will have 1 as a value of both variables, and vehicles that did neither have 0 for both variables (0) No (14734) (1) Yes (15624)
	IMPACT	(0) No Impact (616) (1) Front (14390) (2) Right Side (5417) (3) Left Side (6658) (4) Back (3277)
	ROLLOVER	(0) No rollover (29415) (1) Indicates if a rollover occurred (tripped or untripped). Rollover is defined as any vehicle rotation of 90° or more about any true longitudinal or lateral axis (943)
	MAN_COL	Indicates the orientation of the vehicles in a collision (0) Non-collision accident (5622) (1) Rear-end (3491) (2) Head-on (1601) (3) Angle (19644)
Other information	FRINIGHT	Indicates if the accident happened late Friday night (midnight to 4 a.m. Saturday) (0) No (29926) (1) Yes (432)
	SATNIGHT	Indicates if the accident happened late Saturday night (midnight to 4 a.m. Sunday) (0) No (29911) (1) Yes (447)
	SUNNIGHT	Indicates if the accident happened late Sunday night (midnight to 4 a.m. Monday) (0) No (30227) (1) Yes (131)

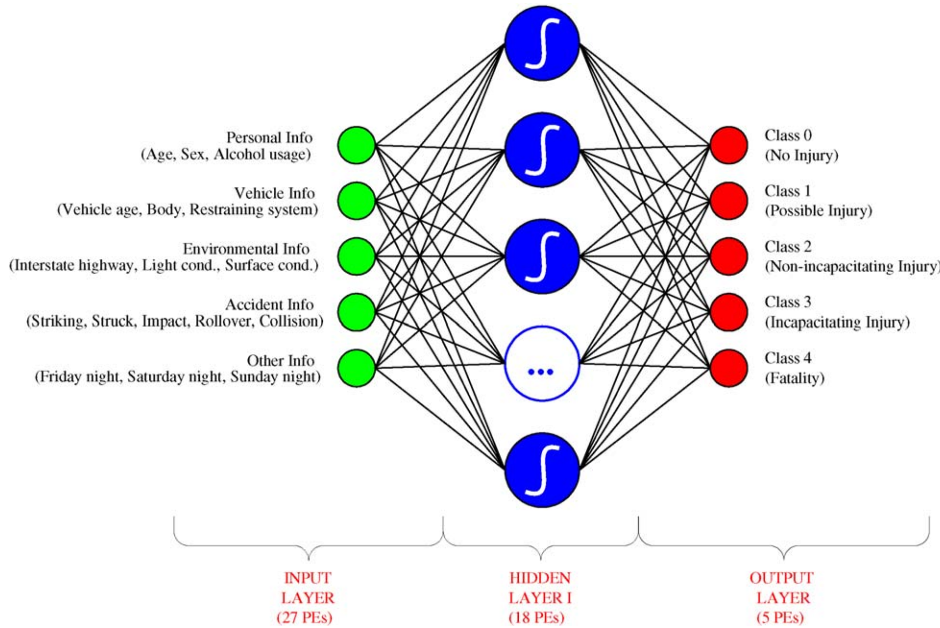

Fig. 1. Graphical representation of our MLP neural network model.

Table 2
Injury severity levels represented in the dataset

Injury severity level	No. of examples	Percent
No injury	10748	35.4
Probable injury	7166	23.6
Non-incapacitating	5941	19.6
Incapacitating	5408	17.8
Fatality	1095	3.6
Total	30358	100.0

acceptable results. As seen in Table 3, a five-class MLP model showed only 40.73% overall prediction accuracy, being able to accurately classify only 8.9% of the fatality cases.

The problem herein can be defined as a multi-class classification problem. In such cases where the data is highly skewed (unbalanced among the class labels) and the problem domain is complex, reducing the multi-class problem into series of two-class (binary) classification problems has been a popular approach that produces better prediction results (Allwein et al., 2000; Tax and Duin, 2002). According to Anand et al. (1995) such conversion in artificial neural networks significantly

Table 3
Prediction accuracy of five-category model (%)

	Training		Testing	
	Mean	S.D.	Mean	S.D.
No injury	53.60	1.89	53.78	2.31
Probable	44.90	18.10	44.73	17.64
Non-incapacitating	53.21	10.46	52.61	10.91
Incapacitating	2.20	3.80	2.60	4.52
Fatal	9.48	19.97	8.86	18.68
Overall	40.71	3.15	40.73	3.17

reduces the model building (training) time while increasing the overall prediction accuracy. Such improvement in ANN model building is credited to the fact that by reducing the multi-class problem into binary classifiers one would increases the utilization of resources (connection weights in artificial neural networks) and simplify the model structure for better and deeper representation. There are two competing approaches in reducing multi-class problems to binary classifiers. First one is called one-vs-all (OVA) where each class is compared to rest of the classes, for an N class problem, resulting in N binary classifiers. Second approach is called all-vs-all (AVA) where every possible combination of individual classes are paired, for an N class problem, resulting in $(N!/((N-2)!*2!))$ binary classifiers. According to Fürnkranz (2002), AVA performance was superior to OVA. However, Rifkin and Klautau (2004) using a wide range of benchmark datasets, showed that simple OVA is as accurate as AVA approach. In this study, after trying a wide range of combinations, we settled on a partitioning schema that is in between OVA and AVA, resulting in eight binary classification models.

Using the above-mentioned method, the complete dataset was separated into eight subsets with binary output variables. Table 4 shows the details of such separation. For better clarity, a graphical representation of the binary model configuration is also given in Fig. 2. The eight partitions are created using a top-down (more serious injury versus the less serious injuries, models 1.1–1.4) and a bottom-up (less serious injury versus more serious injuries, models 2.1–2.4). For each of the eight models, Table 4 shows the number of cases and percent representation of each binary category. Similarly, Fig. 2 illustrates the binary splits for each of the eight models along with their proportional coverage of the complete dataset. A similar partitioning approach is used by Dissanayake and Lu (2002) in identifying the factors influencing severity of injuries to older drivers. After developing a series of binary logistic regression models, they chose to interpret the

Table 4
Series of binary classification models and new datasets composition

Model	Category 0		Category 1	
	Variable	Count (percent)	Variable	Count (percent)
1.1	At most incapacitating	29263 (96.39%)	Fatality	1095 (3.61%)
1.2	At most non-incapacitating	23855 (81.52%)	Incapacitating	5408 (18.48%)
1.3	At most probable	17914 (75.10%)	Non-incapacitating	5941 (24.90%)
1.4	No injury	10748 (60.00%)	Probable	7166 (40.00%)
2.1	No injury	10748 (35.40%)	At least probable	19610 (64.60%)
2.2	Probable	7166 (36.54%)	At least non-incapacitating	12444 (63.46%)
2.3	Non-incapacitating	5941 (47.74%)	At least incapacitating	6503 (52.26%)
2.4	Incapacitating	5408 (83.16%)	Fatality	1095 (16.84%)

coefficients of the "best fit" model as opposed to looking at the problem from a progression of severity perspective (as is the case in our study). Their study revealed that restraint device usage, point of impact, use of alcohol and drugs, gender, travel speed, urban/rural nature and grade/curve existence at the crash location were the important factors for injury severity differences for older drivers involved in single vehicle crashes.

The advantage of the multiple binary models over a single multi-class model is that this approach can provide better insight into the interrelationships between different levels of injury severity and the crash factors. Each model is expected to provide a set of prioritized factors (possibly somewhat different form the other models) specific to a single binary classification, comparing two classes of injury severities, thus offering a deeper insight of what circumstances could be causing increased injury levels during automobile crashes. In addition, some of the binary classification datasets that resulted from this model structure ended up being better-balanced for the classification algorithm.

4.2. k-Fold cross-validation

In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, researchers tend to use k-fold cross-validation (Kohavi, 1995). In k-fold cross-validation, also called rotation estimation, the complete dataset (D) is randomly split into k mutually exclusive subsets (the folds: $D_1, D_2, \ldots, D_k$) of approximately equal size. The classification model is trained and tested k times. Each time ($t \in \{1, 2, \ldots, k\}$), it is trained on all but one folds ($D \backslash D_t$) and tested on the remaining single fold (D_t). The cross-validation estimate of the overall accuracy is calculated as simply the average of the k individual accuracy measures:

$$\text{CVA} = \frac{1}{k}\sum_{i=1}^{k} A_i \tag{2}$$

where CVA stands for cross-validation accuracy, k is the number of folds used, and A is the accuracy measure for each fold.

Since the cross-validation accuracy would depend on the random assignment of the individual cases into k distinct folds, a common practice is to stratify the folds themselves. In stratified k-fold cross-validation, the folds are created in a way so that they contain approximately the same proportion of output class labels (e.g., different levels of injury severities) as the original dataset. Empirical studies show that stratified cross-validation tends to generate comparison results with lower bias and lower variance when compared to regular k-fold cross-validation (Kohavi, 1995).

In this study, we estimate the performance of models and the sensitivity values of independent variables using a stratified 10-fold cross-validation approach. We used 10-fold cross-validation in order to minimize the potential bias associated with the random splitting/sampling of the data into training and testing sets for each of the eight models. The presented results (both model prediction accuracies and variable sensitivity values) in the paper are the averages of the folds. Empirical studies show that 10 seems to be an optimal number of folds (that optimizes the time

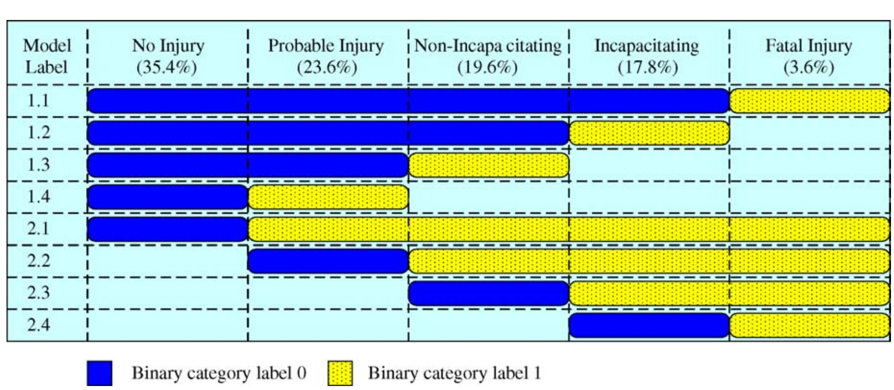

Fig. 2. Graphical representation of binary model configurations.

it takes to complete the test while minimizing the bias and variance associated with the validation process) (Breiman et al., 1984; Kohavi, 1995). In 10-fold cross-validation the entire data set is divided into 10 mutually exclusive subsets (or folds) with approximately the same class distribution as the original data set (stratified). Each fold is used once to test the performance of the classifier (or used to measure the sensitivity values for each independent variables) on the trained neural network model trained on the combined data of the remaining nine folds, leading to 10 independent performance estimates.

4.3. *Sensitivity analysis on the trained neural network models*

Although ANNs are powerful function approximators, they are criticized for being black-box solutions due largely to their inability to explain their results. In response to such criticism, ANN researchers have come up with mathematical models that interpret the ANN weights in order to explain their inference mechanism (Fish and Segall, 2004; Fish and Blodgett, 2003). Sensitivity analysis is one of those promising methods capable of extracting the cause and effect relationship between the inputs and outputs of a trained neural network model (Principe et al., 2000). It has become a commonly used method in neural network studies for identifying the degree to which each input channel (independent variables) contributes to the identification of each output channel (dependent variables). In the process of performing sensitivity analysis, the neural network learning is disabled so that the network weights are not affected. The basic idea is that the inputs to the network are perturbed, and the corresponding change is reported as a percentage deviation in the output. The first input is varied between its mean (median for categorical variables) plus (or minus) a user-defined number of standard deviations (upper and lower value limits for categorical variables), while all other inputs are fixed at their respective means (or medians). The network output is computed and recorded as the absolute percent change above and below the mean of that output variable. This process is repeated for each and every input variable. As an outcome of this process, a report (usually a column plot or a table) is generated that summarizes the variation of each output with respect to the variation in each input.

5. Results and discussion

Neural network function of a popular data mining toolkit, called SPSS Clementine (SPSS, 2005), was used to conduct the experiments. Binary encoding of categorical input variables was used to present the inputs to the neural network. In order to minimize the bias associated with the random selection of training and testing sample sets, a 10-fold cross-validation procedure was employed. Table 5 shows the predictive accuracies of models obtained during the MLP training phase. Table 6 shows the accuracies of models measured on the testing datasets using the trained models. As seen from the results, the binary model structure showed greater accuracy not only for the overall model, which is as high as 90% for some models, but for each of the severity classes compared to the five-category model. Especially significant was the accuracy increase for the prediction of fatality in models 1.1 and 2.4, with accuracy rates of 45.5 and 49.4% as compared with 8.9% correct classification rate provided by the five-class model. While these accuracy rates are not directly comparable because of different model structure, the experiment shows that it is possible to boost the prediction rates of individual categories by restructuring the problem.

Since the focus of this study was on identifying the significant factors influencing levels of injury severity in drivers during car accidents, sensitivity analysis was conducted to identify the most important contributors among the input variables. For this step, the pre-trained network is used to test the sensitivity of the outcome variable to the changes in one of the inputs. This was achieved by varying the values of one of the inputs while keeping all the other inputs fixed at their respective means. This process was then repeated for each input. Finally, a report was generated which summarizes the variation of output with respect to the variation of each input. As the result of the sensitivity analysis, each input variable received a value that is a measure of its relative importance, varying between 0 (meaning it has no effect on the prediction) and 1 (a field that completely determines the prediction). Tables 7 and 8 show the results of sensitivity analysis for all of the binary models. For each of the models, the input variables are listed in the order of their relative importance going from the most to the least significant.

Examination of the sensitivity analysis results reveals considerable differences among the eight models with respect to the significant variables, each having a different set of variables that are the most significant in identifying the injury severity in drivers. For some of these eight models (models 1.1, 1.2, 2.3 and 2.4) a subset of there-to-five variables separates themselves from the rest by having significantly larger sensitivity values (in model 1.1 these significant variables are REST_SYS, AGE, and MAN_COL; in model 1.2 they are REST_SYS, ALC_DRU, ROLLOVER, and AGE; in model 2.3 they are AGE, REST_SYS, INT_HWY, ROLLOVER, and ALC_DRU; and in model 2.4 they are AGE, MAN_COL, and REST_SYS). In other models (models 1.1, 1.2, 2.3 and 2.4) the sensitivity values of input variables do not seem to have an observable separation point between a more significant subset of variables and the rest. An intuitive interpretation of this observation would be that the artificial neural network models designed to differentiate between higher levels of injuries (i.e., fatality incapacitating injury) from the combinations of lower levels of injuries (i.e., non-incapacitating injury and no injury) produce better results than the ones designed to differentiate between lower levels of severity classes.

These sensitivity results also highlight the changing importance levels of predictors among the eight models. As can be seen, the most important variable of a specific model might come out to be unimportant in other models. Additionally, a progression of importance of a variable (or a group of variables) can be observed as the configuration of the models changes from differentiating between more severe injury levels and the lesser ones. For instance, REST_SYS was identified as the most significant predictor in models 1.1 and 1.2. However, it is the least

Table 5
Training model performance (predictive accuracies measured in percent)

	Model 1.1		Model 1.2		Model 1.3		Model 1.4		Model 2.1		Model 2.2		Model 2.3		Model 2.4	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Category 0	90.99	6.57	96.57	1.77	79.95	2.94	89.69	1.10	80.02	0.64	76.41	3.10	64.25	0.29	50.14	3.54
Category 1	46.63	3.09	46.97	4.24	43.46	3.24	68.27	2.10	51.99	0.70	73.36	6.51	99.54	0.58	76.68	7.55
Overall	89.39	6.26	87.41	0.89	70.86	1.52	81.12	0.93	70.10	0.42	75.29	0.75	81.10	0.24	72.21	5.72

Table 6
Testing model performance (predictive accuracies measured in percent)

	Model 1.1		Model 1.2		Model 1.3		Model 1.4		Model 2.1		Model 2.2		Model 2.3		Model 2.4	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Category 0	90.97	6.50	96.48	1.83	79.96	2.28	89.66	1.62	80.00	1.06	75.90	2.65	63.99	0.88	49.37	3.95
Category 1	45.45	3.99	46.36	3.92	43.59	4.41	68.36	1.87	52.09	1.83	74.92	5.69	99.52	0.47	76.16	7.59
Overall	89.34	6.25	87.23	1.01	70.92	1.03	81.15	1.48	70.11	0.90	75.59	1.06	80.98	0.52	71.72	6.24

and second to least significant variable in models 1.3 and 1.4. This provides an interesting insight on the importance of use of restraint systems by drivers. It radically increases in importance with the increasing injury levels.

ROLLOVER was fairly important predictor in all models except model 1.4. This might be interpreted as ROLLOVER being a variable of direct influence on the severe injury levels. The model 1.4 simply distinguishes between no injury and probable injury, while all the other models deal with more serious injury levels. This shows that in the rollover accidents the driver is more likely to get injured. The variable became significant in most models even though there are only 943 cases of rollover accidents in the entire dataset. A quick query of the data revealed that all of these 943 cases resulted in at least some injury to the driver.

According to the sensitivity analysis results, the importance of drivers' gender (the variable named SEX) decreases as being a significant predictor with the increased levels of injury. That is, more serious injuries do not depend on driver being a male or a female. This trend could prompt a more detailed examination of exactly what effects the drivers' sex has on the injury levels sustained during automobile accidents. In a previous study, Abdelwahab and Abdel-Aty (2001) found that while females are more likely to experience a more severe injury, males are more likely to be involved in fatal accidents.

Another interesting observation can be made for the AGE variable. It turns out to be one of the most important predictor for the models that differentiate between high levels of injury severity and other levels of injury severities such as models 1.1, 1.2, 2.3 and 2.4. Such results indicate that AGE is a

Table 7
Relative Importance of Inputs Resulting from Sensitivity Analysis

Model 1.1		Model 1.2		Model 1.3		Model 1.4	
REST_SYS	0.2028	REST_SYS	0.5387	ROLLOVER	0.2542	SEX	0.3980
AGE	0.1863	ALC_DRU	0.4548	ALC_DRU	0.1791	BODY_TYP	0.2986
MAN_COL	0.1735	ROLLOVER	0.4385	BODY_TYP	0.1763	STRIKING	0.2163
SUNNIGHT	0.0440	AGE	0.3879	STRIKING	0.1576	AGE	0.1977
ROLLOVER	0.0415	BODY_TYP	0.0836	SEX	0.1390	STRUCK	0.1671
INT_HWY	0.0388	SEX	0.0640	INT_HWY	0.1104	IMPACT	0.1305
IMPACT	0.0272	STRIKING	0.0582	IMPACT	0.0953	MAN_COL	0.1261
STRUCK	0.0214	INT_HWY	0.0528	SUNNIGHT	0.0773	SUNNIGHT	0.0865
BODY_TYP	0.0114	IMPACT	0.0478	FRINIGHT	0.0747	FRINIGHT	0.0681
SATNIGHT	0.0109	STRUCK	0.0333	AGE	0.0737	INT_HWY	0.0598
ALC_DRU	0.0104	MAN_COL	0.0299	MAN_COL	0.0691	VEH_AGE	0.0510
STRIKING	0.0065	SUR_COND	0.0232	STRUCK	0.0559	SUR_COND	0.0326
VEH_AGE	0.0064	LGHT_CON	0.0198	VEH_AGE	0.0399	LGHT_CON	0.0302
SEX	0.0053	VEH_AGE	0.0175	SUR_COND	0.0355	SATNIGHT	0.0107
LGHT_CON	0.0047	SUNNIGHT	0.0143	LGHT_CON	0.0321	ALC_DRU	0.0102
FRINIGHT	0.0036	FRINIGHT	0.0129	SATNIGHT	0.0261	REST_SYS	0.0043
SUR_COND	0.0034	SATNIGHT	0.0052	REST_SYS	0.0100	ROLLOVER	0.0036

REST_SYS: restraining system; AGE: age of driver; MAN_COL: orientation of vehicle in collation; SUNNIGHT: Sunday night; ROLLOVER: rollover occurred; INT_HWY: interstate highway; IMPACT: direction of impact; STRUCK: crash role; BODY_TYP: vehicle body type; SATNIGHT: Saturday night; ALC_DRU: under influence; STRIKING: crash role; VEH_AGE: vehicle age; SEX: sex of driver; LGHT_CON: light conditions; FRINIGHT: Friday night; SUR_COND: surrounding conditions.

Table 8
Relative importance of inputs resulting from sensitivity analysis

Model 2.1		Model 2.2		Model 2.3		Model 2.4	
BODY_TYP	0.2957	REST_SYS	0.2931	AGE	0.5627	AGE	0.3874
SEX	0.2813	ROLLOVER	0.2754	REST_SYS	0.5348	MAN_COL	0.3432
ROLLOVER	0.2379	IMPACT	0.2727	INT_HWY	0.4895	REST_SYS	0.2081
REST_SYS	0.2369	ALC_DRU	0.2721	ROLLOVER	0.4343	INT_HWY	0.1658
ALC_DRU	0.2284	VEH_AGE	0.2539	ALC_DRU	0.4121	SUNNIGHT	0.1604
STRIKING	0.1829	STRUCK	0.1898	LGHT_CON	0.0694	IMPACT	0.1248
AGE	0.1774	STRIKING	0.1423	SEX	0.0509	ROLLOVER	0.0956
IMPACT	0.1521	MAN_COL	0.1375	SUR_COND	0.0445	STRUCK	0.0503
STRUCK	0.1153	AGE	0.1279	IMPACT	0.0140	BODY_TYP	0.0420
SUNNIGHT	0.1081	INT_HWY	0.0798	SUNNIGHT	0.0118	SATNIGHT	0.0342
MAN_COL	0.0996	SEX	0.0354	FRINIGHT	0.0115	SEX	0.0314
FRINIGHT	0.0764	SUR_COND	0.0309	BODY_TYP	0.0104	LGHT_CON	0.0263
INT_HWY	0.0591	BODY_TYP	0.0283	SATNIGHT	0.0082	STRIKING	0.0227
LGHT_CON	0.0411	FRINIGHT	0.0265	MAN_COL	0.0056	FRINIGHT	0.0215
SUR_COND	0.0404	SUNNIGHT	0.0210	STRIKING	0.0052	ALC_DRU	0.0134
VEH_AGE	0.0097	SATNIGHT	0.0139	STRUCK	0.0036	VEH_AGE	0.0130
SATNIGHT	0.0091	LGHT_CON	0.0099	VEH_AGE	0.0029	SUR_COND	0.0122

REST_SYS: restraining system; AGE: age of driver; MAN_COL: orientation of vehicle in collation; SUNNIGHT: Sunday night; ROLLOVER: rollover occurred; INT_HWY: interstate highway; IMPACT: direction of impact; STRUCK: crash role; BODY_TYP: vehicle body type; SATNIGHT: Saturday night; ALC_DRU: under influence; STRIKING: crash role; VEH_AGE: vehicle age; SEX: sex of driver; LGHT_CON: light conditions; FRINIGHT: Friday night; SUR_COND: surrounding conditions.

significant predictor for injury severity of drivers for serious automobile crashes. Similar results were found by Dissanayake and Lu (2002) where they study the factors influencing severity of injuries to older drivers using a series of linear regression models.

Examination of the sensitivity results for the STRIKING and STRUCK variables shows that the importance of STRIKING decreases with the increased injury severity levels while the importance of STRUCK increases. While the fact that a vehicle is striking some object or another vehicle is significant, it is not a good predictor of increasing the probability of a more severe injury. However, the fact that a vehicle is being struck by another is more effective in increasing the probability of a more serious injury level. This might be due to a fact that an occupant of a striking vehicle might have more time to react to the situation and prepare for impact then an occupant of a struck vehicle.

BODY_TYP variable seems to show an increasing trend of importance as the model type changes from 1.1 through 1.4. Remember that this variable denotes the type of vehicle involved in the accident such as passenger car, sport utility vehicle, van and pickup truck. Model 1.4 differentiates between no injury and probable injury accidents where the BODY_TYP seem to be a significant predictor (second most important one after the SEX variable) whereas model 1.1 differentiates between fatality and all other injury levels including no injury where BODY_TYP seem to be one of the least important variables (ninth overall). It is difficult to interpret such results. One might argue that even though such phenomenon with BODY_TYP variable cannot be explained singularly, it may very well be explained in the context of changes with other variables. Another set of experiments where the model sensitivity measures are to be calculated not only for single variables but also for between effects of multiple variable combinations may be warranted.

6. Conclusions

This paper has presented a study for identification of person, vehicle, and accident characteristics influential in making a difference in injury severity levels sustained by drivers in traffic accidents. Eight binary MLP neural network models were developed with different levels of injury severity as the dependent variable. These eight models presented different levels of injury severity varying from the no-injury to fatality and from fatality to no-injury. All models were found to have better predictive power as compared to a model with a five-category outcome variable. In addition, this structure helped to identify the important explanatory variables at each level of distinction between the injury severities. The use of a restraint system like a seat belt, use of alcohol or drugs, persons' age and gender, and vehicle role in the accident were found to have an important influence on the outcome of the crash. At the same time, weather conditions or the time of the accident did not seem to affect the severity risk of injury. This result seemed a bit surprising and needs further study. As suggested by one of the reviewers, drivers may adjust their driving style (or not drive altogether) according to their perception of the risk of bad weather. It is also possible that some accident types may be not adequately represented because, for example, certain weather conditions are rather rare. It is also possible that a non-linear mapping technique such as a neural network model is better able to accommodate the possibly cumulative, non-linear effects of multiple characteristics of a situation leading to different accident injury levels. No single factor by itself appears to be a key determinant of accident severity, but can act as a catalyst or a barrier in combination with other factors in affecting the injury severity levels. These results have implications for policy makers, transportation system designers, and researchers. Transportation safety designers cannot easily identify a single factor, make recommendations for incremental changes in that factor, and hope to achieve major differences in injury severity levels. The problems have to be analyzed and attacked from a multidimensional perspective: vehicle characteristics, alcohol, road characteristics, collision avoidance systems, etc. Researchers similarly may adopt techniques such as neural networks for analysis of such variables.

The results presented in this study validate some of the findings of previous studies, but with a much larger dataset and with a method that achieves higher predictive accuracy by analyzing the generalized mapping techniques of neural modeling. These also provide insight into the changing importance of crash factors with the changing injury severity levels. The results also highlight cases where further research is necessary to come up with better understanding of changing behavior of certain crash factors.

Acknowledgement

We acknowledge the invaluable help and insight provided by Mr. Vish Daita and Mr. Paul Bobowski from Ford Motor Company.

References

Abdel-aty, M.A., Chen, C., Schott, J.R., 1998. An assessment of the effect of driver age on traffic accident involvement using log-linear models. Acc. Anal. Prev. 30 (6), 851–861.

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalizes Intersection. Transportation Research Record 1746, Paper No. 01-2234, pp. 6–13.

Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Acc. Anal. Prev. 34 (6), 729–741.

Allwein, E.L., Schapire, R.E., Singer, Y., 2000. Reducing multiclass to binary: a unifying approach for margin classifiers. J. Mach. Learn. Res. 1, 113–141.

Anand, R., Mehrotra, K., Mohan, C.K., Ranka, S., 1995. Efficient classification for multiclass problems using modular neural networks. IEEE Trans. Neural Networks 6 (1), 117–124.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Dissanayake, S., Lu, J.J., 2002. Factors influential in making an injury severity difference to older drivers involved in highway crashes. Acc. Anal. Prev. 34 (5), 609–618.

Edwards, J., 1998. The relationship between road accident severity and recorded weather. J. Safety Res. 29 (4), 249–262.

Farmer, C.M., Braver, E.R., Mitter, E.L., 1997. Two-vehicle side impact crashes: the relationship of vehicle and crash characteristics to injury severity. Acc. Anal. Prev. 29 (3), 399–406.

Fish, K.E., Blodgett, J.G., 2003. A visual method for determining variable importance in an artificial neural network model: an empirical benchmark study. J. Target. Meas. Anal. Market. 11 (3), 244–254.

Fish, K.E., Segall, R.S., 2004. A visual analysis of learning rule effects and variable importance for neural networks employed in data mining operations. Kybernetes: Int. J. Syst. Cybernet. 33 (5/6), 1127–1142.

Fürnkranz, J., 2002. Round robin classification. J. Mach. Learn. Res. 2, 721–747.

GES, 2005. National Automotive Sampling System General Estimates System, National Center for Statistics and Analysis of the National Highway Traffic Safety Administration.

Haykin, S., 1998. Neural Networks: A Comprehensive Foundation. Prentice-Hall, New Jersey.

Hornik, K., Stinchcombe, M., White, H., 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward network. Neural Networks 3, 359–366.

Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: an application of ordered probit models. Acc. Anal. Prev. 34 (3), 313–321.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Wermter, S., Riloff, E., Scheler, G. (Eds.), The Fourteenth International Joint Conference on Artificial Intelligence (IJCAI). Montreal, Quebec, Canada. Morgan Kaufman Publishing, San Francisco, CA, pp. 1137–1145.

Lui, K.J., McGee, D., 1988. An application of conditional logistic regression to study the effects of safety belts, the principal impact points, and car weights on drivers' fatalities. J. Safety Res. 19 (4), 197–203.

Mussone, L., Ferrari, A., Oneta, M., 1999. An analysis of urban collisions using an artificial intelligence model. Acc. Anal. Prev. 31, 705–718.

O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Acc. Anal. Prev. 28 (6), 739–753.

Principe, J.C., Euliano, N.R., Lefebre, W.C., 2000. Neural and Adaptive Systems: Fundamentals Through Simulations. John Wiley and Sons, New York.

Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. J. Mach. Learn. Res. 5, 101–141.

SPSS, 2005. SPSS Clementine for Windows. SPSS Inc., Chicago, IL.

Tax, D.M.J., Duin, R.P.W., 2002. Using two-class classifiers for multiclass classification. In: Proceedings of the 16th International Conference on Pattern Recognitions, vol. 2, August 11–15. IEEE Press, pp. 124–127.

Wilson, R., Sharda, R., 1994. Bankruptcy prediction using neural networks. Decis. Support Syst. 11, 545–557.

Wood, D.P., Simms, C.K., 2002. Car size and injury risk: a model for injury risk in frontal collisions. Acc. Anal. Prev. 34, 93–99.