Accident Analysis and Prevention 61 (2013) 3–9

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Individual driver risk assessment using naturalistic driving data

Feng Guo[a,*], Youjia Fang[b]

[a] Department of Statistics, Virginia Tech Transportation Institute, Virginia Tech, 406A Hutcheson Hall, Blacksburg, VA 24061-0439, USA
[b] Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA

ARTICLE INFO

Article history:
Received 30 November 2011
Received in revised form 6 June 2012
Accepted 18 June 2012

Keywords:
Individual driver risk
Naturalistic Driving Study
NEO-5 Personality inventory
Critical incident
K-mean cluster

ABSTRACT

Driving risk varies substantially among drivers. Identifying and predicting high-risk drivers will greatly benefit the development of proactive driver education programs and safety countermeasures. The objective of this study is twofold: (1) to identify factors associated with individual driver risk and (2) predict high-risk drivers using demographic, personality, and driving characteristic data. The 100-Car Naturalistic Driving Study was used for methodology development and application. A negative binomial regression model was adopted to identify significant risk factors. The results indicated that the driver's age, personality, and critical incident rate had significant impacts on crash and near-crash risk. For the second objective, drivers were classified into three risk groups based on crash and near-crash rate using a *K*-mean cluster method. The cluster analysis identified approximately 6% of drivers as high-risk drivers, with average crash and near-crash (CNC) rate of 3.95 per 1000 miles traveled, 12% of drivers as moderate-risk drivers (average CNC rate = 1.75), and 84% of drivers as low-risk drivers (average CNC rate = 0.39). Two logistic models were developed to predict the high- and moderate-risk drivers. Both models showed high predictive powers with area under the curve values of 0.938 and 0.930 for the receiver operating characteristic curves. This study concluded that crash and near-crash risk for individual drivers is associated with critical incident rate, demographic, and personality characteristics. Furthermore, the critical incident rate is an effective predictor for high-risk drivers.

1. Introduction

The substantial variation in individual driving risk has been documented in many studies (Deery and Fildes, 1999; Ulleberg, 2001; Dingus et al., 2006). Identifying factors associated with individual driving risk and predicting high-risk drivers will enable proper driver-behavior intervention and safety countermeasures to reduce the crash likelihood of high-risk groups and improve overall driving safety.

Traffic safety research involves drivers, vehicles and driving environment. There are extensive literatures on the safety impact of transportation infrastructure and traffic characteristics, e.g., the impacts of intersection design features, pavement conditions, weather, and traffic flow conditions (Hauer et al., 1988; Poch and Mannering, 1996; Maze et al., 2006; Guo et al., 2010; Lord and Mannering, 2010). Crash occurrence is the primary risk measure for infrastructure-related safety impact evaluation, with Poisson and negative binomial (NB) models being the state-of-practice analysis tools. However, there are limited researches on individual driver risk in traffic and human factor engineering fields.

Contrary to traffic engineers, the insurance and actuarial science industries have a long history of research on classification of drivers according to risk level to facilitate underwriting and pricing. Estimation of the occurrence of claims based on the driver's age and other relevant variables has been a standard practice in actuarial research (Segovia-Gonzalez et al., 2009). For the insurance industry, quantified individual risk is directly related to the risk classification standards (Walters, 1981). However, insurance data are proprietary and, in general, not available for public access.

Individual driver risk can be affected by many factors. Besides demographic variables such as age and gender, driver personality – commonly measured by the NEO five traits inventory or Zuckerman's Sensation Seeking Scale, – also plays an important role in individual driving risk (Costa and McCrea, 1992). Studies have shown the association between personality characteristics and risky driving behavior (Jonah, 1997; Jonah et al., 2001; Ulleberg and Rundmo, 2003; Dahlen and White, 2006; Machin and Sankey, 2008).

Driver behavior plays a central role in driver risk but it is difficult to measure in real-world driving situations. Recent developments in vehicle instrumentation techniques, such as in Naturalistic Driving Study (NDS) (University of Michigan Transportation Research Institute, 2005; Dingus et al., 2006; Guo and Hankey, 2009) and the DriveCam system (Hickman et al., 2010) have made it both technologically possible and economically feasible to monitor driving

* Corresponding author. Tel.: +1 540 231 1038; fax: +1 540 231 3863.
E-mail addresses: feng.guo@vt.edu (F. Guo), youjia@vt.edu (Y. Fang).

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.06.014

behaviors and kinematic signatures on a large scale. These data collected through advanced in-vehicle instrumentation provide an opportunity to link the driver behavior with risk at the individual driver level.

NDSs collect rich kinematic, Global Positioning System (GPS), radar, and video data at a high frequency, which provides an opportunity to detect abnormal driving situations. In particular, the authors are interested in whether critical-incident events (CIEs) – non-crash safety events marked by a high acceleration/deceleration rate or other kinematic signatures – can be used to predict high-risk drivers. The premise is that critical incidents are caused by driver behaviors similar to that of CNCs. Since critical incidents happen at a much higher frequency (100 times the frequency of crashes and 10 times the frequency of near-crashes), this provides an opportunity to identify high-risk drivers before accidents actually happen. This will allow designing and implementing proactive safety countermeasures to improve the safety of the high-risk drivers.

The objectives of this study are twofold. The first objective is to investigate the risk factors associated with individual driving risk. The second objective is to build up a model to predict high-risk drivers, which includes two steps: identification using cluster analysis, and prediction using a logistic regression model. The 100-Car Naturalistic Driving Study was used for methodology development and application.

2. Materials and methods

2.1. The 100-Car Naturalistic Driving Study data

The 100-Car Naturalistic Driving Study is the first large-scale NDS conducted in the United States (Dingus et al., 2006). The study included 102 primary drivers in northern Virginia. In order to catch as much safety critical events as possible, the samples lean towards young drivers and high mileage drivers. The vehicles of the participants were instrumented with advanced data acquisition systems. The system included five camera views (forward, driver face, over the shoulder, left and right mirror), GPS, speedometer, three-dimension accelerometer, and radar, etc. Driving data were collected continuously for 12 months. The study collected data for approximately 2,000,000 vehicle miles and almost 43,000 h of data.

The data were reduced based on the kinematic and video records. Three types of safety-related events were identified: crashes, near-crashes, and safety-critical events (Dingus et al., 2006; Klauer et al., 2006). A crash is defined as an event with "any contact between the subject vehicle and another vehicle, fixed object, pedestrian, pedacyclist, or animal" (Dingus et al., 2006, p. xvii). The crash involves kinetic energy transfer or dissipation. A near-crash is "a conflict situation that requires a rapid, severe evasive maneuver to avoid a crash. The rapid, evasive maneuver involves conducting maneuvers that involve steering, braking, accelerating, or any combination of control inputs that approaches the limits of the vehicle capabilities" (Dingus et al., 2006, p. xvii).

The CIE is a conflict less severe than the near-crash. CIEs were detected by three approaches (Dingus et al., 2006): (1) flagging events where the car sensors exceeded a specified value (e.g., brake response of >0.6 g); (2) when the driver pressed an incident push-button located on the data acquisition system; (3) through analysts' judgments when reviewing the video. A rigorious data reduction was implemented by using different threshold values for the kinematic threshold values and visual confirmation.

Although not a safety concern by itself, the CIE can be regarded as a measure of driving aggressiveness. The hypothesis is that a relatively safe driver, based on his/her driving skills and safety consciousness, will try to avoid evasive maneuvers that could lead to a hazardous scenario, including a CIE. A high rate of CIEs reflects the lack of such skills and safety consciousness; thus, the rate of CIEs is an indicator of driving aggressiveness. If the above hypothesis holds, the rate of CIEs will be a good predictor for individual driver risk.

Other factors that may be associated with different driving risks include age, gender, and personality. The 100-Car Naturalistic Driving Study included a survey that measures personalities based on the NEO Five-Factor Inventory, which includes the following five aspects: Neuroticism (N), Extroversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C) (Costa and McCrea, 1992; Klauer et al., 2006). A number of research studies have been conducted to evaluate the relationship between the NEO five factors with driving safety (Shaw and Sichel, 1971; Loo, 1979; Arthur and Graziano, 1996; Klauer et al., 2006).

Due to the relatively small number of crashes, near-crashes are commonly used as a crash surrogate. Several research studies in risk assessment using NDS used near-crashes in conjunction with crashes for risk assessment (Klauer et al., 2006; Guo et al., 2010; Klauer et al., 2010). Guo et al. (2010) concluded that the near-crash is a valid crash surrogate for risk assessment purposes. Based on the research cited above, a combination of crash and near-crash events was used as a risk metric for individual driving risk.

2.2. Statistical methods

The study was designed to evaluate two objectives: assess risk factors and predict high-risk drivers. For the first objective, a state-of-the-practice negative binomial (NB) model was used to assess the relationship between the CNC risk and potential risk factors. For the second objective, there are two steps for the prediction of high-risk drivers. First, a *K*-mean cluster analysis was used to identify high-risk driver groups. Logistic regression models were then developed to predict the high-risk drivers using the risk factors identified in the first objective. The prediction performance of the logistic regression model was evaluated by the receiver operating characteristics curve (ROC). The details of the models and the analysis techniques are discussed in this section.

2.2.1. Negative binomial model for evaluating risk factors (Objective 1)

The NB regression model is state-of-the-practice for traffic safety modeling (Lord and Mannering, 2010). The model assumes that the observed frequency of crashes and near-crashes for driver *i*, Y_i, follows an NB distribution:

$$Y_i \sim NB(E_i\lambda_i, \gamma)$$

where λ_i is the expected CNC rate for driver *i*, as measured by the number of CNCs per 1000 miles; E_i is the miles traveled by driver *i* (per 1000 miles); and γ is the NB over-dispersion parameter. A log link function connects λ_i with a set of covariates:

$$\log(\lambda_i) = \mathbf{X_i}\boldsymbol{\beta}$$

where $\mathbf{X_i}$ is the matrix of covariates for driver *i* and $\boldsymbol{\beta}$ is the vector of regression parameters. In this study, the age, gender, and personality score based on the NEO five-factor inventory, and the critical incident were used as covariates.

2.2.2. Cluster analysis for identifying high-risk drivers (Objective 2, Step 1)

The main criterion for evaluating the overall risk of individual drivers is the CNC rate. The cluster analysis provides an objective approach to classify drivers into different risk levels and has been used in traffic safety research (Donmez et al., 2010). A *K*-mean cluster method was adopted to classify primary drivers into different risk groups based on CNC rate. The *K*-mean cluster partitions the

Table 1
Summary statistics by age and gender.

Variables	Age <25		Age 25–55		Age >55	
	Male	Female	Male	Female	Male	Female
Number of drivers	16	18	39	16	8	5
Total number of CIEs	1234	2209	2490	930	490	41
Total number of CNCs	163	224	174	105	61	8
Subject miles (KMiles)	160.7	204.2	525.2	142.9	105.2	192.0
Mean CIE rate[a]	8.2	11.37	4.861	7.63	4.57	2.579
Mean CNC rate[a]	1.11	1.27	0.38	0.73	0.58	1.10
Mean CIE rate[a]	9.88		5.67		3.81	
Mean CNC rate[a]	1.20		0.48		0.78	

[a] Unit of rate is number of events per 1000 miles traveled.

observations into k clusters with a predetermined number of clusters (Tan et al., 2005). An observation is assigned to the cluster whose mean is closest to its value. The K-mean method minimizes the within-cluster sum of squares:

$$\underset{s}{\arg\min} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| X_j - \mu_i \right\|^2$$

where $(X_1, X_2, \ldots, X_n)$ are the observed data which are the CNC rates in the context of this paper; $\boldsymbol{S} = (S_1, \ldots, S_k)$ is the set of k clusters; and μ_i is the mean of the observations in set S_i.

Each driver was classified into one of three clusters (high-, moderate-, and low-risk groups). Drivers in the clusters with the highest mean CNC rate were considered to be high-risk drivers.

2.2.3. Logistic regression models for predicting high-risk drivers (Objective 2, Step 2)

After risk groups were identified through cluster analysis, two logistic regression models were developed to model the probability of being a high-risk driver. The first model evaluates the probability of high-risk drivers only, while the second model evaluates the probability of high- or moderate-risk drivers. The two models could support the interest of researchers with different perspectives. The model setup is as follows. Define

$$Y_i = \begin{cases} 1 & \text{If driver } i \text{ is a high risk driver (or a hig/moderate risk driver)} \\ 0 & \text{Otherwise} \end{cases}$$

Let p_i be the probability of being a risky driver for drive i. The observed Y_i is assumed to follow a Bernoulli distribution.

$$Y_i \sim Bernoulli(p_i)$$

The key parameter is the probability of being a high/moderate risk driver, p_i. This probability is associated with a set of covariates by a logit link function,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{X_i}\boldsymbol{\beta}$$

where $\mathbf{X_i}$ is the matrix of predictors for individual i, and $\boldsymbol{\beta}$ is the vector of regression parameters. The exponential of regression parameter, $\exp(\beta_j)$, is the odds ratio (OR) for the j^{th} variable. The CIE rate, age group, and personality score were used as driver characteristics. The logistic regression will estimate the probability of being a risky driver based on predictors. A driver will be predicted as a risky driver if this probability is greater than a predefined threshold value p_0.

The predictive performance of the logistic models was evaluated by the ROC curve (Agresti, 2002), which measures model sensitivity and specificity. In the context of this study, the sensitivity is the probability of correctly predicting a risky driver, and the specificity is the probability of correctly predicting a safe driver, as shown in the following formula, i.e.,

Sensitivity = Probability (Classified as risky driver | the driver is risky)

Specificity = Probability (Classified as safe driver | the drive is safe)

Both measures were related to the threshold value p_0 and there is a tradeoff between sensitivity and specificity. The ROC curve is a plot of sensitivity versus false positive rate; i.e., (1 – Specificity), for all possible thresholds p_0's. The performance of the prediction model can be measured by the area under the curve (AUC): a higher AUC value indicates better prediction power for the logistic regression model. A perfect prediction method would yield the maximum AUC of 1. A completely random guess would give a diagonal line in the ROC space with AUC of 0.5.

3. Results

3.1. Exploratory data analysis

The 100-Car Study data include 60 crashes, 675 near-crashes, and 7394 critical incidents from primary drivers. The event rate was calculated as number of events per 1000 miles traveled:

$$\text{Event Rate} = \frac{\text{Number of Events}}{\text{Miles Travelled (1000 miles)}}$$

Based on overall risk by age and sample size considerations, three age groups were defined: younger than 25 years, between 25 and 55 years, and older than 55 years. The summary statistics stratified by age and gender are shown in Table 1.

Drivers under the age of 25 had the highest CIE and CNC rates among all the age groups. Drivers between 25 and 55 had a higher CIE rate than did drivers older than 55 but had a lower CNC rate. The CNC and CIE rates also vary by gender and age group. Male drivers have lower CIE and CNC rates than female drivers in the <25 and 25–55 age groups. The gender difference is not consistent for drivers older than 55. Male drivers over 55 had a lower CNC rate but a higher CIE rate as compared to female drivers over 55.

The NEO five-factor personality inventory represents various aspects of personality using five variables. The data analysis indicated that the Extroversion, Agreeableness, and Conscientiousness factors have strong correlations with the response CNC rate, as shown in Table 2. However, the five variables themselves are highly correlated. Including all factors in the same model will lead to multicollinearity issues and biased inference. Choosing a subset of variables could mitigate the multicollinearity issues but would lead to insufficient use of information. The principal component analysis (PCA) was adopted to address the multicollinearity and maintain the maximum information from the five variables.

The PCA uses an orthogonal transformation to convert correlated variables into a set of uncorrelated variables called principal

Table 2
Pearson correlation coefficients between Neo-5 personality scores and response.

	N[a]	O[a]	E[a]	A[a]	C[a]	CIE rate	CNC rate
N	1.00	0.64[b] <0.0001[c]	0.62 <0.0001	0.59 <0.0001	0.35 0.0004	0.03 0.7851	−0.17 0.0955
O		1.00	0.62 <0.0001	0.58 <0.0001	0.43 <0.0001	0.03 0.7762	−0.12 0.2049
E			1.00	0.70 <0.0001	0.63 <0.0001	−0.13 0.2111	−0.20 0.0473
A				1.00	0.65 <0.0001	−0.20 0.0456	−0.26 0.0093
C					1.00	−0.14 0.1706	−0.21 0.0381

[a] N: neuroticism; O: openness to experience; E: extroversion; A: agreeableness; C: conscientiousness.
[b] Correction coefficients.
[c] *P*-value, two sided test under the null hypothesis of zero correlation.

Table 3
Eigenvalues for principal components.

Component	Eigenvalue	Difference	Proportion	Cumulative
1	3.365	2.679	0.673	0.673
2	0.686	0.311	0.137	0.810
3	0.374	0.067	0.074	0.885
4	0.306	0.03	0.061	0.946
5	0.267	–	0.053	1.000

components (Jolliffe, 2002). The first principal component has the highest variance and accounts for the largest portion of the variability in the data. Each succeeding component in turn has the highest variance possible under the constraint of orthogonality (uncorrelated) with the preceding components. A principal component is a linear combination of optimally weighted observed variables.

The first step of the principal component analysis is to identify significant principal components for the set of correlated observed variables. The eigenvalue-one criterion was used to choose the significant component, which states that a component contains substantial information if the corresponding eigenvalue is greater than 1. As can be seen from Table 3, the eigenvalue of the first component is 3.365 (much larger than 1) and all four of the other components have eigenvalues smaller than 1. The first component could contribute to 67.3% of the variability in the data. Therefore, the first component from the PCA was used to represent the personality scores. A sensitivity analysis was also conducted using the first two components and the results indicate essentially identical prediction power. Therefore, one component is considered sufficient in risk modeling.The results of the PCA provide the following formula for the personality score:

$$\text{PCA personality score}_i = 0.232 \times N_i^* + 0.261 \times E_i^* + 0.237 \times O_i^* + 0.256 \times A_i^* + 0.231 \times C_i^*$$

where the N_i^* to C_i^* are standardized values for driver *i*. For example, $N_i^* = N_i - \bar{N}/std.dev.(N)$, with the $\overline{N}$ and $std.dev.(N)$ being the mean and standard deviation of the observed variable *N*, respectively.

3.2. Negative binomial models for risk factors evaluation

The NB regression model was fitted using the CNC frequency as the response variable and the miles traveled as exposure (per 1000 miles). Three covariates were included: the CIE rate as measured by the number of CIEs per thousand miles, the age group, and the personality score as computed by PCA. The gender variable does not have significant impacts on the CNC rate and was excluded from the NB model and subsequent analysis. The model fitting results are shown in Table 4. As can be seen, all three factors are highly significant. The over-dispersion parameter is quite small (0.262), which indicated the presence of over-dispersion and justified the use of NB regression. The point estimate for the critical incident parameter is 0.091, which implies that for every one unit of increase in critical incident rate, the CNC rate will increase to a multiplicative factor of $exp(0.091) = 1.09$. That is an approximately 10% increase in CNC rate for every one unit of increase in CIE rate.

The 25–55 age group showed the lowest crash and near-crash rates. The CNC rate ratio between drivers younger than 25 years and drivers between 25 and 55 years is $exp(0.541) = 1.72$; for drivers older than 55 and drivers between 25 and 55 years it is $exp(0.481) = 1.62$.

3.3. Identify driver risk groups

The *K*-mean cluster method was applied to the 102 primary drivers based on CNC rate. The number of clusters was predefined to be three, to represent the high, moderate, and low driver risk groups. The number of clusters is determined based on sample size and ease of interpretation for the modelling results.

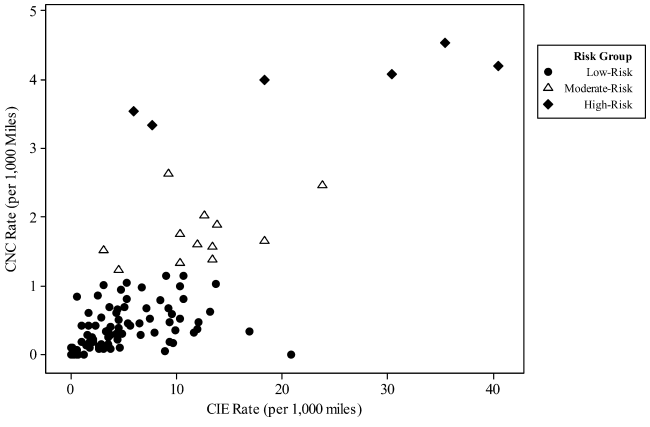

Fig. 1. Cluster results.

Table 4
Parameter estimation for negative binomial models.

Parameters	Estimate	Standard error	Wald 95% Confidence limits		*P*-value
Intercept	−1.557	0.125	−1.802	1.312	<0.0001
CI rate	0.091	0.011	0.070	0.112	<0.0001
Age: <25 vs. 25–55	0.541	0.157	0.234	0.848	0.0006
Age: >55 vs. 25–55	0.481	0.234	0.021	0.940	0.0402
Personality	−0.317	0.082	−0.477	−0.157	0.0001
Dispersion	0.262	0.069	0.141	0.424	

Table 5
Characteristics of driver risk groups.

Risk groups	Number of drivers	Mean CNC rate	% of males in each group	Mean age	Means of the NEO personality factors				
					A	E	O	N	C
Low-risk group	84	0.39	65.5	38.1	38.0	36.7	35.1	25.4	37.2
Moderate-risk group	12	1.75	58.3	28.7	33.3	33.5	31.7	22.5	32.2
High-risk group	6	3.95	16.7	30.0	29.7	31.7	34.7	21.3	32.0

Table 6
Logistic regression model output.

Models	Effect	Parameter estimate	*P*-Value	Odds ratio	95% Odds ratio confidence limits	
Model 1: high vs. moderate/low risk	Intercept	−11.1	0.023	–	–	–
	CIE	**0.346**	**0.019**	**1.414**	**1.058**	**1.889**
	Age: <25 *vs.* 25–55	4.37	0.175	79.42	0.142	>999
	Age: >55 *vs.* 25–55	6.46	0.108	638.2	0.242	>999
	Personality	−0.900	0.432	0.407	0.043	3.838
Model 2: high/moderate vs. low risk	Intercept	−5.62	<0.0001	–	–	–
	CIE	**0.296**	**<0.001**	**1.345**	**1.150**	**1.571**
	Age: <25 *vs.* 25–55	**2.13**	**0.021**	**8.456**	**1.378**	**51.90**
	Age: >55 *vs.* 25–55	2.36	0.059	10.64	0.918	123.3
	Personality	−0.48	0.368	0.616	0.214	1.770

The output of the cluster analysis is illustrated in Fig. 1. A relatively small number of drivers were in the risk groups (6 drivers in the high-risk group and 12 drivers in the moderate-risk group). As the goal of the study is to identify risky drivers, the relatively small numbers in these two groups fit the context well. The within-cluster variations for the low-, moderate-, and high-risk groups were 0.31, 0.44, and 0.44, respectively.

The characteristics of the three risk groups are summarized in Table 5. As can be seen, the CNC rate of the high-risk group is 10 times that of the safe (i.e., low-risk) group, and the rate of the moderate-risk group is more than 4 times that of the safe group. The average age of the safe group is substantially higher (38.1). The overall pattern of the NEO personality factors suggests that the low-risk group has relatively high values in all five factors and the high-risk group has relatively low values in the NEO five-factors.

3.4. Logistic prediction models results for predicting high-risk drivers

The cluster analysis identified the high-, moderate-, and low-risk groups. The key question is whether the high-risk drivers can be predicted by the driver characteristics. Depending on specific research questions, it could be of interest to predict extremely high-risk drivers or moderate- to high-risk drivers. Therefore, two logistic prediction models were developed. The first model predicted high-risk drivers against moderate-risk/safe drivers. The second model predicted high/moderate-risk drivers against the safe drivers. The risk factors identified in the NB regression model were used in logistic regression; i.e. the critical incident rate, age group, and PCA component based on the NEO-5 personality score. The model outputs are summarized in Table 6.

In both models the CIE rate had a significant impact on the probability of being a risky driver. The OR was calculated to quantitatively evaluate the impact s of each variable. The OR represents the relative odds of being a risky driver for every one unit increase in a continuous variable (critical incident rate and personality score), or relative risk between two levels of a categorical covariate (the age group variable). The results indicated that, for every one unit increase in CIE rate, the relative odds of being a high-risk driver will increase by 41% (OR = 1.414). Based on Model 2, every one unit increase in CIE rate will increase the relative odds of being a moderate/high-risk driver by 35% (OR = 1.345).

The personality score variable, differing from the NB models, did not show significant results in the logistic regression models. The age group variables are not significant in the prediction model for high-risk drivers. However, the model for predicting moderate/high-risk drivers indicates a significant difference between the young driver group (<25) and the middle age group (25–55). One potential cause for the discrepancy between factors identified in the NB model and the logistic regression model is that the cluster process masks the CNC rate difference among drivers within the same group.

The predictive models are as follows,

Model 1:

$$\text{Probability (high-risk driver)} = \frac{\exp(-11.1 + 0.346 \times \textit{CIE Rate} + 4.37 \times \textit{AgeL25} + 6.46 \times \textit{AgeG55} - 0.900 \times \textit{PER})}{1 + \exp(-11.1 + 0.346 \times \textit{CIE Rate} + 4.37 \times \textit{AgeL25} + 6.46 \times \textit{AgeG55} - 0.900 \times \textit{PER})}$$

Model 2:

$$\text{Probability (high- or moderate-risk driver)} = \frac{\exp(-5.62 + 0.296 \times \textit{CIE Rate} + 2.13 \times \textit{AgeL25} + 2.36 \times \textit{AgeG55} - 0.48 \times \textit{PER})}{1 + \exp(-5.62 + 0.296 \times \textit{CIE Rate} + 2.13 \times \textit{AgeL25} + 2.36 \times \textit{AgeG55} - 0.48 \times \textit{PER})}$$

where *PER* is the standardized personality score, *AgeL25* is a indicator variable on whether the driver age is less than 25; and *AgeG55* is a indicator variable on whether driver age is greater than 55.

To evaluate the prediction performance, ROCs for both models were generated as shown in Fig. 2. The solid bold lines are the ROC curves, and the straight diagonal dashed lines are the reference lines. Both models showed high predictive power. The AUC is 0.938 for Model 1 and 0.930 for Model 2, both close to the perfect AUC value of 1.

4. Discussion

The NDS collects rich real-life driving data for an extended period of time. The continuous data collection approach provides the opportunity to evaluate not only crash risk, but also non-crash driving behavior. This study evaluated the individual driving risk and the risk factors associated with high-risk drivers using the 100-Car Naturalistic Driving Study data, the first large-scale NDS conducted in the United States.

Driving risk varies substantially among drivers. The cluster analysis indicated that about 6% of the drivers had a substantially higher risk (10 times higher than low-risk groups) and about 12% of the drivers showed moderate to high risk (4 times higher than the low-risk group). This result is consistent with the substantial variation in driving risk observed from previous NDS, epidemiological, self-reporting, and simulator studies (Deery and Fildes, 1999; Ulleberg, 2001; Dingus et al., 2006; Donmez et al., 2010). Although high-risk drivers only account for a small proportion of the driver population, they have a substantial impact on overall traffic safety. The ability to identify high-risk drivers will provide a valuable reference for developing safety education programs, regulations, and proactive safety countermeasures.

The NB regression model indicated that driver personality, age, and CIE rate had significant impacts on the CNC risk for individual drivers. It is well known that young and elderly drivers have a higher risk as compared to other age groups (National Highway Traffic Safety Administration, 2008). The relative risk between age groups from this analysis is consistent with the national data.

The relationship between driver personality and driving risk has been evaluated in previous studies (Ulleberg and Rundmo, 2003; Dahlen and White, 2006; Machin and Sankey, 2008). It is surprising to observe high correlations among the five factors, which were designed to measure different aspects of personality. The PCA was used to utilize maximum information from all five factors without inducing the multicollinearity issue. This study confirmed that the NEO five-factor inventory did associate with CNC risk for individual drivers.

A primary focus of this study is whether CIE rate, a measure of driving characteristics, can be used to assess driving risk and predict high/moderate-risk drivers. The results confirmed that the CIE rate has a strong relationship with individual driving risk. Furthermore, a logistic prediction model using critical incident rates can successfully identify high- and moderate-risk drivers. The CIE rate had a statistically significant influence on the prediction. For every one-unit increase in CIE rate, the relative probability of being a high-risk or moderate-to-high risk driver increased by approximately 41% and 35%, respectively.

The strong association between the CIE rate and individual driver risk can have significant implications on the individual risk assessment and safety interventions. Since the number of accidents for individual drivers is often limited, predicting risky drivers using past accident history may be inefficient. The CIEs occur at a much higher frequency than crashes (one hundredfold higher) and near-crashes (tenfold higher). This makes it possible to proactively identify the high-risk population. This is particularly important for developing proactive safety countermeasures and for improving the safety of high-risk drivers. There have been studies that not only monitored driver behaviors but also attempted to improve safety by providing feedback to alter driver behaviors (Donmez et al., 2010). The insurance industry has also begun incorporating driving characteristics into pricing; e.g., the SnapShot® program by the Progressive Casualty Insurance Company.

Traffic accidents are rare events, thus, surrogates are needed when there is not a sufficient number of accidents for safety assessment (Tarko et al., 2009). Traffic conflict is one of the most widely used surrogates (Williams, 1981; Hauer and Garder, 1986; Tiwari

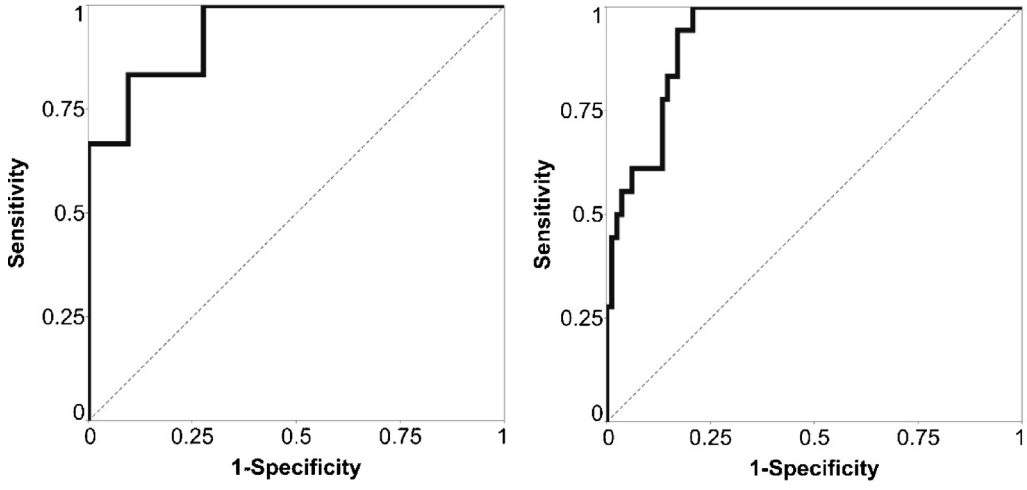

Fig. 2. The ROC curves.

et al., 1998). Surrogates are especially critical for NDSs which usually observe a limited number of crashes but with high resolution. Guo et al. (2010) concluded that the near-crash could provide valuable information on driving risk and can serve as a crash surrogate for risk-assessment purposes, especially for time-variant risk factors such as distraction. The strong association between the CIE rate and CNC risk at the individual driver level, as well as the strong prediction power for high-risk drivers, implies that CIEs could be a valid safety surrogate for driver risk analyses.

Although this study is limited by the voluntary and relatively small sample size, a significant relationship was observed between individual driving risk and factors such as demographic, personality, and driving characteristics. More importantly, it demonstrates that naturalistic data can provide opportunities to identify individual driver risk differences from less severe safety events, that when observed over time can provide a more complete profile of unsafe driving behavior. This study is a first step and additional studies with larger and more representative data are clearly needed. Large scale naturalistic studies currently underway in the US (Committee for the Strategic Highway Research Program 2: Implementation, 2009) and in other countries, such as the EuroFOT program (Benmimoun et al., 2009), hold promise for providing the data needed to expand upon the research described in this paper. The outcomes of this research can help designers, educators, and researchers developing better proactive safety countermeasures and safety programs to improve the driving safety.

5. Conclusion

Using the 100-Car Naturalistic Driving Study data, this study showed that individual drivers' driving risk varies substantially with three distinct risk groups. The cluster analysis identified approximately 6% of drivers as high-risk drivers, which accounted for 24% of total crash and near-crash events; and 12% of drivers as high/moderate-risk drivers, which accounted for 45% of total crash and near-crash events. The NB regression model indicated that age, personality, and CIE rate had significant impacts on individual drivers' CNC risk. The logistic prediction models for high- and moderate-risk groups had high predictive powers using the CIE rate. This study concluded that CNC risk for individual drivers is associated with CIE rate, age, and personality characteristics. Furthermore, the CIE rate is an effective predictor for high-risk drivers.

References

Agresti, A., 2002. Categorical Data Analysis, 2nd ed. Wiley-Interscience, New York.

Arthur, W., Graziano, W.G., 1996. The five-factor model, conscientiousness, and driving accident involvement. Journal of Personality 64 (3), 593–618.

Benmimoun, A., Benmimoun, M., Van Noort, M., Wilmink, I., 2009. Eurofot: Large scale field operational test – impact assessment. In: 16th ITS World Congress, Stockholm.

Committee for the Strategic Highway Research Program 2: Implementation, 2009. Implementing the results of the second strategic highway research program: Saving lives, reducing congestion, improving quality of life - special report 296. The National Academies Press.

Costa, P.T., McCrea, R.R., 1992. Revised NEO personality inventory (NEO PI-R) and NEO Five-Factor Inventory. Psychological Assessment Resources, Odessa, Fl.

Dahlen, E.R., White, R.P., 2006. The big five factors, sensation seeking, and driving anger in the prediction of unsafe driving. Personality and Individual Differences 41 (5), 903–915.

Deery, H.A., Fildes, B.N., 1999. Young novice driver subtypes: relationship to high-risk behavior, traffic accident record, and simulator driving performance. Human Factors: The Journal of the Human Factors and Ergonomics Society 41 (4), 628–643.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen A., Lee S.E., Sudweeks J.D., Perez, M.A., Hankey J.M., Ramsey D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jermeland, J., and Knipling, R.R. 2006. The 100-car naturalistic driving study: phase II – Results of the 100-car field experiment. Report No.: DOT HS 810 593. National Highway Traffic Safety Administration, Washington, DC.

Donmez, B., Boyle, L.N., Lee, J.D., 2010. Differences in off-road glances: effects on young drivers' performance. Journal of Transportation Engineering-Asce 136 (5), 403–409.

Guo, F., Hankey, J.M., 2009. Modeling 100-Car Safety Events: A case-based approach for analyzing naturalistic driving data. the National Surface Transportation Safety Center for Excellence.

Guo, F., Klauer, S.G., Hankey, J.M., Dingus, T.A., 2010. Using near-crashes as a crash surrogate for naturalistic driving studies. Transportation Research Record: Journal of the Transportation Research Board 2147, 66–74.

Hauer, E., Garder, P., 1986. Research into the validity of the traffic conflicts technique. Accident Analysis & Prevention 18 (6), 471–481.

Hauer, E., Ng, J., Lovell, J., 1988. Estimation of safety at signalized intersections. Transportation Research Record: Journal of the Transportation Research Board, 48–61.

Hickman, J.S., Hanowski, R.J., Bocanegra, J., 2010. Distraction in commercial trucks and buses: assessing prevalence and risk in conjunction with crashes and near-crashes. Report No. FMCSA-RRR-10-049. Federal Motor Carrier Safety Administration, Washington, DC.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer.

Jonah, B.A., 1997. Sensation seeking and risky driving: a review and synthesis of the literature. Accident Analysis & Prevention 29 (5), 651–665.

Jonah, B.A., Thiessen, R., Au-Yeung, E., 2001. Sensation seeking, risky driving and behavioral adaptation. Accident Analysis & Prevention 33 (5), 679–684.

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J., 2006. The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data. Report No. DOT HS 810 594. National Highway Traffic Safety Administration, Washington, DC.

Klauer S.G., Guo F., Sudweeks J.D., Dingus T.A., 2010. An analysis of driver Inattention using a case-crossover approach on 100-Car data. Report No. DOT HS 811 334. National Highway Traffic Safety Administration, Washington, DC.

Loo, R., 1979. Role of primary personality factors in the perception of traffic signs and driver violations and accidents. Accident Analysis & Prevention 11 (2), 125–127.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291–305.

Machin, M.A., Sankey, K.S., 2008. Relationships between young drivers' personality characteristics, risk perceptions, and driving behaviour. Accident Analysis & Prevention 40 (2), 541–547.

Maze, T.H., Agarwai, M., Burchett, G., 2006. Whether weather matters to traffic demand traffic safety, and traffic operations and flow. Transportation Research Record: Journal of the Transportation Research Board 1948, 170–176.

National Highway Traffic Safety Administration, 2008. Traffic safety facts 2006: A compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system.

Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-accident frequencies. Journal of Transportation Engineering 122, 105–113.

Segovia-Gonzalez, M.M., Guerrero, F.M., Herranz, P., 2009. Explaining functional principal component analysis to actuarial science with an example on vehicle insurance. Insurance: Mathematics and Economics 45, 278–285.

Shaw, L., Sichel, H.S., 1971. Accident Proneness: Research in the Occurrence, Causation and Prevention of Road Accidents. Pergamon, Oxford, England.

Tan, P.N., Steinbach, M., Kumar, V., 2005. Introduction to Data Mining. Addison-Wesley.

Tarko, A., Davis, G., Saunier, N., Sayed, T., Washington, S., 2009. Surrogate measure of safety: White paper. Transportation Research Board ANB20(3) Subcommittee on Surrogate Measures of Safety.

Tiwari, G., Mohan, D., Fazio, J., 1998. Conflict analysis for prediction of fatal crash locations in mixed traffic streams. Accident Analysis & Prevention 30 (2), 207–215.

Ulleberg, P., 2001. Personality subtypes of young drivers. Relationship to risk-taking preferences, accident involvement, and response to a traffic safety campaign. Transportation Research Part F: Traffic Psychology and Behaviour 4 (4), 279–297.

Ulleberg, P., Rundmo, T., 2003. Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. Safety Science 41 (5), 427–443.

University of Michigan Transportation Research Institute, 2005. Automotive collision avoidance system field operational test methodology and results appendices. National Highway Traffic Safety Administration.

Walters, M.A., 1981. Risk classification standards. Proceedings of the Casualty Actuarial Society 68, 1–18.

Williams, M.J., 1981. Validity of the traffic conflicts technique. Accident Analysis & Prevention 13 (2), 133–145.