INDIVIDUAL DRIVER RISK ANALYSIS USING NATURALISTIC DRIVING DATA

Feng Guo, Ph.D
Assistant Professor, Department of Statistics,
Virginia Tech Transportation Institute, Virginia Tech,
Blacksburg, VA, USA, email: feng.guo@vt.edu

Youjia Fang
Research Assistant, Department of Statistics,
Virginia Tech, Blacksburg, VA, USA, email: youjia@vt.edu

Submitted to the 3rd International Conference on Road Safety and Simulation, September 14-16, 2011, Indianapolis, USA

ABSTRACT

Individual driver risk varies substantially and a small percentage of drivers often contribute to a disproportionate large number of safety events. To identify factors associated with individual drivers and predict high-risk drivers will benefit the development of driver education programs and safety counter-measures. The goal of this study is twofold 1) to assess risk factors associated individual drivers and 2) to predict high-risk drivers. The 100-Car Naturalistic Driving Study data was used for methodology development. A negative binomial regression analysis indicated that driver age, extroversion of the NEO 5 personality trait, and critical incidents, as a measure of aggressive driving behavior, had significantly 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. Approximate 6% of drivers were identified as high-risk and 18% of driver as high/moderate risk drivers. A logistic regression model was developed to predict high risk drivers as well as high/moderate risk drivers. The predictive models showed high predicting power with area under the curve value of 0.917 and 0.9351 for the receiver operating characteristic curves. This study concluded that age, personality, and driving behavior is closely related to individual driving risk and aggressive driving is a powerful predictor for high risky drivers.

Keywords: Individual Driver Risk, Prediction, Naturalistic Driving Study, NEO personality test, Critical Incident

INTRODUCTION

The "good driver and bad driver" phenomenon, i.e., a relative small number of drivers contribute to an disproportionate large number of safety events, has been documented in many studies (Dingus et al 2006 Deery et al 1999, Ulleberg, 2001). In the 100-Car Naturalistic Driving Study (NDS), 10 percent of drivers contributed to near 35 percent of all crashes and near-crashes (Dingus et al 2006). To identify factors associated with individual driving risk and predict high-risk drivers will allow using proper driver behavior intervention or safety countermeasures to reduce the crash likelihood of the high risk groups thus improve the overall traffic safety.

Individual driver risk is not a focus of traditional traffic safety engineering research in general. Instead, engineers are more interested in safety impacts of transportation infrastructure and traffic characteristics, e.g., the impacts of intersection design features, pavement conditions, weather, and traffic flow conditions on traffic safety with Poisson and negative binomial model being the primary modeling tools (Guo *et al* 2010, Hauer, Ng and Lovell 1988, Maze, Agarwai and Burchett 2006, Poch and Mannering 1996, Lord and Mannering, 2010).

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

Individual driver risk can be affected by many factors. Beside demographic variables such as age and gender, driver personality, commonly measured by NEO 5 traits test, also plays and important role in individual driving risk (Costa and McCrae1992). Studies have shown the association between risky driving behavior and personality characteristics (Ulleberg and Rundmo, 2003, Dahlen and White, 2006, Machin and Sankey, 2008).

Driver behavior plays a central role in individual driver risk but is difficult to measure in real driving situations. Recent development in vehicle instrumentation techniques, such as the naturalistic driving study (Dingus *et al*, 2006, Guo and Hankey 2009) and DriveCam system

(Hickman et al, 2010) have made it both technologically possible and economically feasible to monitoring the driving behaviors and kinematic signatures as well as driver behavior at large scale. There has been studies not only monitored the driver behavior but also attempted to improve safety by providing feedback to alter the driver behavior. These data provide an opportunity to link the driver behavior with risk at individual driver level.

Naturalistic driving studies collected high frequency data, which allow us to detect abnormal driving situations. In particular, we are interested in whether critical incident (CI), a non-crash safety event marked by high acceleration/deceleration rate or other kinematic metrics, can be used to predict high risk drivers. The premise is that critical incidents are caused by driver behaviors similar to that of crash and near-crash. Since critical incidents happen at much high frequency (100 times of crash frequency and 10 times of near-crash frequency), it provide an opportunity to identify high risk drivers before accidents actually happen. This will provide an opportunity for design and implement proactive safety countermeasures to improve the safety of those drivers.

Traffic accidents are rare events and surrogates need to be used when there are no sufficient number of accidents for safety assessment (Tarko et al 2009). Traffic conflicts is one of the most widely used surrogates (Tiwari et al 1998, Willimans 1981, Hauer and Garder 1986, Williman, 1981). Surrogates are especially critical for naturalistic driving studies which usually observe limited number of crashes but high resolution driving data. Guo et al (2010a) conducted the first systematic attempt to address the crash surrogate measure for naturalistic driving study. The results indicated that the near-crash could provide valuable information on driving risk and can serve as a crash surrogate for risk assessment purpose.

The objectives of this study is twofold. The first objective is to investigate risk factors associated with individual driving risk. The second objective is to predict high risk drivers, which includes two steps: identification and prediction of high risk drivers. The 100-Car Naturalistic Driving Study, the first large scale naturalistic driving study, was used for methodology development.

THE 100-CAR NATURALISTIC DRIVING 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. The

vehicles of the participants were instrumented with advanced data acquisition system. The system included five camera views (forward, driver face, over the shoulder, left and right mirror), GPS, speedometer, three-dimension accelerometer, 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 hours of data. The study included 102 primary drivers and more than 100 secondary drivers.

The data were reduced based on the kinematic and videos records. Three type of safety related events were identified: crash, near-crash, and safety critical event (Dingus et al 2006, Klauer et al 2006). The crash is defined as an event where "any contact between the subject vehicle and another vehicle, fixed object, pedestrian pedacyclist, or animal." The crash evolves kinetic energy transfer or dissipation. The 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 involves steering, braking, accelerating, or any combination of control inputs that approaches the limits of the vehicle capabilities".

The critical incident is also a conflict but is less severe than the near-crash. Critical incidents 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 pushbutton located on the data acquisition system; and 3) through analysts' judgments when reviewing the video.

The critical incident is an abnormal driving event. However, itself does not directly provide useful information in assessing the safety impacts of driver distraction and other potential risk factors (Klauer et al 2006). In this study, we treated the critical incident as a measure of the driving aggressiveness of individual drivers. The hypothesis is that for a relative safe driver, based on his/her driving skill and safety consciousness, will try to avoid evasive maneuvers that could lead to a hazardous scenario. A high rate of critical incident reflects the lack of such skills and safety consciousness. Thus the rate of critical incident is an indicator of driving aggressiveness. If the above hypothesis holds, the rate of critical incident will be a good predictor for individual driver risk.

Other factors that are could lead to difference driving risks include age, gender and personalities. The 100-Car Study included a survey that measures personalities based on the

NEO Five-Factor Inventory, which includes the following five aspects: the Neuroticism (N), Extroversion (E), Openness to Experience (O), Agreeableness (A) and Conscientiousness (C) (Costa and McCrae, 1992, Klauer et al 2006). A number of researches have been conducted to evaluate the relationship between the NEO five factors with driving safety (Klauer et al 2006, Arthur and Graziano,1996;; Loo, 1979; and Shaw and Sichel, 1971).

Due to the relative small number of crashes, near-crash is commonly used as a crash surrogate. Several key researches in risk assessment using naturalistic driving study used near-crash in conjunction with crash for risk assessment (Klauer et al 2006, Klauer et al 2010, Guo and Hankey, 2010). Guo et al (2010) evaluated the appropriateness and consequence of using near-crashes as crash surrogates. Their study indicated that crash and near-crash shares similar causal mechanism and there is a strong frequency relationship between crashes and near-crashes. Based on the above researches, we used both crash and near-crash (CNC) as a safety metric for individual driving risk.

METHODOLOGY

The study was designed to evaluate both objectives, i.e. assess risk factor and predict high risk drivers, as illustrated in Figure 1. A negative binomial model was used to assess risk factors. The high risk driver predictive model consists of two steps. The first step is to identify high risk drivers thorough a K-mean cluster method. The second step is to predict high risk drivers as identified in the first step. A logistic regression model as used to build the prediction model. The performance of the prediction is evaluated by the receiver operation curve (ROC). The details for each model is discuss in this session.

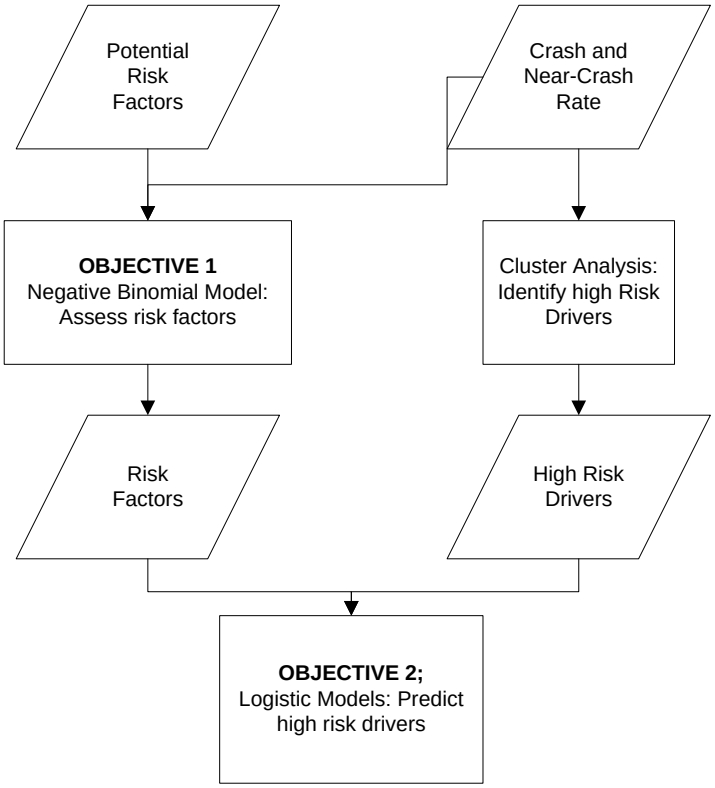


Figure 1 Study Structure

Negative Binomial Model for Evaluating Risk Factors (Objective 1)

The NB regression model is the state-of-the-practice for traffic safety modeling (Lord and Mannering, 2010). The model assumes the observed frequency of crash and near-crash for driver i, Y_i, follows a negative binomial distribution,

$$y_i \sim NB(E_i\lambda_i, r)$$

where λ_i is the expected CNC rate for driver i, as measured by number of CNC per mile, E_i is the miles traveled by driver i, and r is an over-dispersion parameter. A log link function collects λ_i with a set of covariates,

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

where $\boldsymbol{X_i}$ is the matrix of covariates for driver i and $\boldsymbol{\beta}$ is the vector of regression parameters. In this study the age, gender, NEO personality factors, and critical incident were used as covariates.

Cluster Analysis for Identifying High Risk Drivers (Object 2 Step 1)

A K-mean cluster method was used to classify primary drivers into difference risk groups based on CNC rate. The K-mean cluster partitions the observations into *k* clusters with predetermined number of clusters (Tan *et al* 2005). An observation is assigned to the cluster whose means is most close to its values. The K-mean method minimize the within-cluster sum of squares

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|X_j - \mu_i\|^2$$

where $(X_1, X_2, \dots, x_n)$ is the observed data which is the CNC rate in the context of this paper; $\boldsymbol{S} = \{S_1, \dots, S_k\)$ is the set of k clusters, and μ_i is the mean of the observations in set S_i.
Each driver will classified into one of the clusters. Drivers in the clusters with the highest mean CNC rate were considered as high risky drivers.

Logistic Regression Models for Predicting High Risk Drivers (Object 2 Step 2)

A logistic regression model was developed to model the probability of being a risky driver, as identified thorough cluster analysis, based on a set of covariates. The model setup is as follows. Define

$$\mathrm{Y_i} = \begin{cases} 1 & \text{If driver } i \text{ is an high risk driver.} \\ 0 & \text{Otherwise} \end{cases}$$

Let p_i be the probability of being a risky driver for driver i. The observed Y_i is assumed to follows a Bernoulli distribution which is a binomial distribution with sample size one.

$$Y_i \sim Bernoulli\ (p_i)$$

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

$$\mathrm{logit}(p_i) = \log(\frac{p_i}{1-p_i}) = \boldsymbol{X_i \beta}$$

where $\boldsymbol{X_i}$ 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 estimated odds ratio for the j^{th} variable. Again the critical incident rate, age group, and NEO personality factors were used as predictors. The logistic regression will generate the probability of being a risky driver based on predictors. A driver was predicted as a risky driver if this probability is greater than a threshold value p_0.

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

$$Sensitivity = probability(Classified\ as\ risky\ driver|the\ drive\ is\ risky)$$
$$Specificity = probability(Classified\ as\ safe\ driver|the\ drive\ is\ safe)$$

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 vs. (1 – Specificity) for all possible threshold p_0 . The performance of the prediction model can be measure by the area under the curve (AUC), the higher (closer to 1) the AUC is, the better the prediction power for that logistic regression model. The best possible prediction method would yield an AUC of 1, representing 100% sensitivity and 100% specificity. A completely random guess would give a diagonal line of the ROC space with AUC 0.5. .

RESULTS

Exploratory Data Analysis

The 100 Car Study data include 60 crashes, 675 near-crashes, and 7,394 critical incidents from primary drivers. The event rate was calculated as number of event per 1000 miles traveled:

$$Event\ Rate = \frac{\text{Number of events}}{1000\ \text{Miles traveled}}$$

Three age groups were defined : < 25, 25-55, and > 55. This classification was based on overall risk by age and sample size considerations. Other classification was also examined, e.g., <20, 20-40, 40-60, >60, but produced poor model fitting. The summary statistics stratified by age and gender are shown in Table 1.

Table 1 Summary Statistics by Age and Gender

	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 CI	1234	2209	2490	930	490	41
Total Number of CNC	163	224	174	105	61	8
Subject Miles (KMiles)	160.7	204.2	525.2	142.9	105.2	192.0
Mean CI rate*	8.2	11.37	4.861	7.63	4.57	2.579
Mean CNC rate*	1.11	1.27	0.38	0.73	0.58	1.10
Mean CI rate*	9.88		5.67		3.81	
Mean CNC rate*	1.20		0.48		0.78	

*the unit of rate is number of events per 1000 miles traveled

For age group younger than 25 years old, both event rates are consistently the highest among age groups. Age group 25 to 55 has a higher critical incident rate than age group older than 55 but a lower CNC rate. Male drivers has lower critical rate and CNC rate than female drivers in age group <25 and 25-55, while the trend is not clear in age group >55.

The correlation among the predictors and response are listed in Table 2. As can be seen the NEO five factors E, A and C have a very good correlation with the response crash and near-crash rate. However, the factors are also highly correlated with each other. Including all factors in the same model will lead to multi-collinearity problems and biased inference. Using the principle component analysis could alleviate this issue but requires more sophisticated modeling and suffers difficulties in interpretation. We adopted a relative simple approach by only including factor E, which has a high correlated with the response CNC rate but with relative low correlation with another predictor, the critical incident rate.

The extraversion measures not only sociability but also assertiveness, general optimism and cheerfulness. People who score lower on this scale are not pessimists but rather prefer solitude, are generally more subdued in expressing emotion and demonstrate higher levels of cynicism (Costa and McCrae, 1992)

Table 2 Pearson Correlation Coefficients between Predictors and Response

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0							
	N	**O**	**E**	**A**	**C**	**CI Rate***	**CNC Rate***
N	1.00	0.64 <.0001	0.62 <.0001	0.59 <.0001	0.35 0.0004	0.03 0.7851	-0.17 0.0955
O		1.00	0.62 <.0001	0.58 <.0001	0.43 <.0001	0.03 0.7762	-0.12 0.2049
E			1.00	0.70 <.0001	0.63 <.0001	-0.13 0.2111	**-0.20** **0.0473**
A				1.00	0.65 <.0001	**-0.20** **0.0456**	**-0.26** **0.0093**
C					1.00	-0.14 0.1706	**-0.21** **0.0381**

Objective 1: Negative Binomial Model Results for Evaluating Risk Factors

The NB regression model was fitted using CNC frequency as response variable and mileage as exposure. Three covariates were included, the critical incident rate as measured by number of events per thousand miles, the age group, and the extroversion variable. The model fitting results were shown in Table 3. As can be seen, all three factors are highly significantly. The over dispersion parameter is quite small (0.282), which indicated the presence of over-dispersion and justified the use of negative binomial regression. The point estimate for the critical incident parameter is 0.089, which implies that for every one unit of increase in critical incident rate, the CNC rate will increase by a multiplicative factor of $exp(0.089) = 1.09$. That is approximate 10% increase in crash and near-crash rate for every one unit of increase in CI rate.

The extroversion factor coefficient is a negative value of -0.025. With the one unit increase in the extroversion factor, the CNC rate will decrease by a factor of $\exp(-0.025) = 0.975$. The age group 25-55 showed the lowest crash and near crash rate. The CNC rate ratio between age group 25-55 and >55 is $\exp(-0.539) = 0.58$ and between age group 25-55 and <25 is $\exp(-0.65 - 0.065) = 0.49$.

Table 3 Parameter estimation for negative binomial models

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		P-Value
Intercept	-0.142	0.314	-0.757	0.474	0.6523
CI Rate	0.089	0.011	0.067	0.111	<.0001
Age: 25-55 vs. >55	-0.539	0.240	-1.008	-0.069	0.0245
Age: <25 vs. >55	0.065	0.252	-0.429	0.559	0.7954
Extroversion	-0.025	0.007	-0.038	-0.012	0.0002
Dispersion	0.282	0.072	0.141	0.424	

Objective 2 Step 1: Identify Risk Drivers

The K-mean cluster method was applied to the 102 drivers based on CNC rate. The number of clusters was predefined to three to represent high, moderate, and low risk groups. There are several reasons to select three clusters. First, the parsimony principle and interoperability of the results limits the number of clusters. Second, because the CNC rate is bounded by zero, we expect the variations of CNC rate among safe drivers will be small and one cluster was expected to be sufficient for the safe drivers. The CNC rate for risky driver, based on the exploratory analysis, dispersed widely. As the K-mean cluster method is sensitive for the large values, it was expected that one cluster is not sufficient to include all risky drivers. The cluster outputs were consistent this hypothesis.

The output of the clusters analysis was illustrated in Figure 2. Relative small number drivers were in the risky 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 high risky drivers, the relative small sample in these two groups fits the context well. The within cluster variations were 0.31, 0.44, 0.44 respectively.

The characteristics of the three risk groups are summarized in Table 4. As can be seen the CNC rate of the high risk group is 10 times that of the safe group and the rate of moderate – risky group is more than 4 times of the safe group.

Interestingly the percentage of male is the lowest in the high-risky group, only 16.7%. The average age of the safe group is substantially higher (38.1). The overall pattern of the NEO personality factor seems suggests that low risk group has relative high values in all five factors and high risk group has relative low values in all five groups

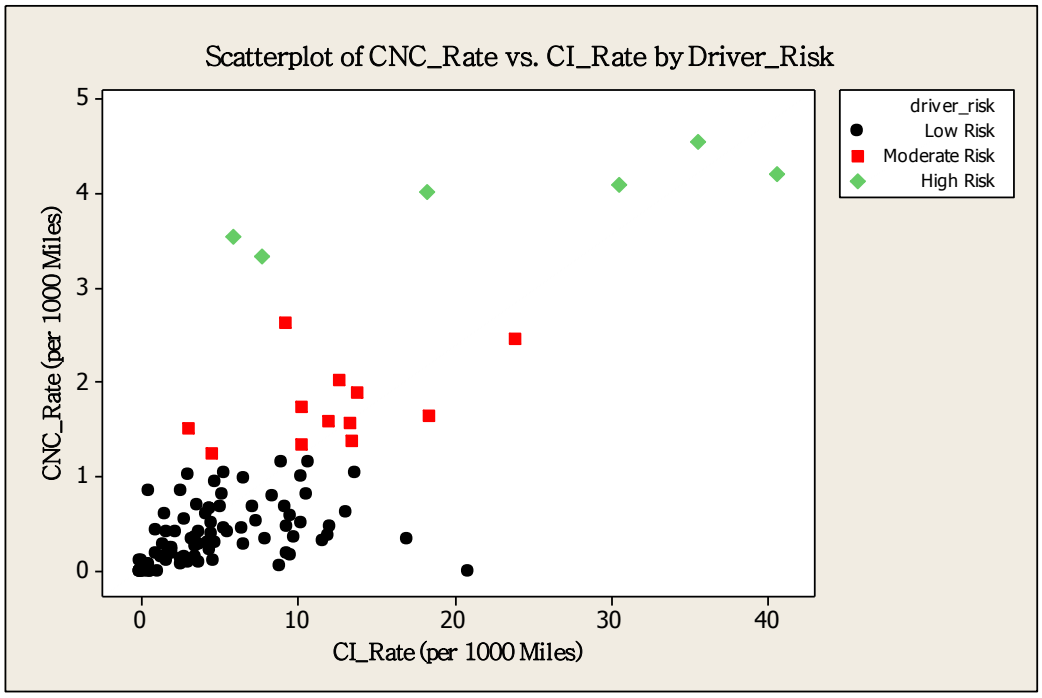


Figure 2 Cluster Results

Table 4 Characteristic of Driver Risk Groups

	# of drivers	Mean CNC rate	% of male drivers	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

- Neuroticism (N), Extroversion (E), Openness to Experience (O), Agreeableness (A) and Conscientiousness (C)

Objective 2 Step 2: Logistic Predict Models Results for Predict High Risk Drivers

Based on the cluster analysis, two high-risky groups and one safe group were identified. Depending on specific research questions, it could be of interest to predict extreme high risky drivers or moderate to high risky drivers. Therefore two logistic prediction models were developed to predict high risk drivers and high/moderate drivers respectively.

The first model predicted high-risk drivers against moderate-risky and safety drivers. The second model predicted high and moderate risk drivers against the safety drivers. The same covariates used in the NB regression model were also used in logistic regression, i.e., the critical

incident rate, age group and extroversion of the NEO five factors. The model outputs were summarized in the Table 5.

In both models the critical incident rate had a significant impact on the probability of being a risky driver. The odds ratio (OR) was calculate to quantitatively evaluate the impacts of each variable (

Table 6). The ORrepresents the relative odds of being a risky driver for every one unit increase in a continuous variable (critical incident rate and extroversion), or relative risk between two levels of a categorical covariate, (the age group variable). As, for every one unit of increase in incident rate, the relative odds of being a high-risky will increase by 40% (OR=1.375). For Model 2, every one unit increase in incident rate will increase the relative odds of being a moderate to high risky drier by 30% (OR=1.311).

Table 5 Logistic Regression Model Parameter Estimate

	Parameter	Estimate	Standard Error	Pr > ChiSq
Model 1: High vs. Moderate/safe	**Intercept**	-3.07	2.0636	0.1366
	Critical Incident Rate	0.32	0.133	**0.0154**
	Age: 25-55 vs. >55	-5.86	3.6033	0.1039
	Age: <25 vs. >55	-2.00	1.6276	0.2192
	Extroversion	-0.033	0.06	0.5563
Model 2 High/Moderate vs. safe	**Intercept**	-1.93	1.35	0.1551
	Critical Incident Rate	0.27	0.07	**0.0002**
	Age: 25-55 vs. >55	-2.29	1.21	0.0588
	Age: <25 vs. >55	-0.22	0.99	0.8243
	Extroversion	-0.032	0.03	0.3358

Table 6 Odds Ratio Estimate from the Logistic Regression Model

	Odds Ratio Estimates			
	Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Model 1: High vs. Moderate/safe	**Critical Incident Rate**	**1.379**	**1.063**	**1.790**
	Age: 25-55 vs. >55	0.003	<0.001	3.330
	Age: <25 vs. >55	0.135	0.006	3.288
	Extroversion	0.968	0.867	1.080
Model 2 High/Moderate vs. safe	**Critical Incident Rate**	**1.311**	**1.137**	**1.511**
	Age: 25-55 vs. >55	0.101	0.009	1.089
	Age: <25 vs. >55	0.802	0.114	5.624
	Extroversion	0.969	0.909	1.033

The extroversion and age group variables, differ from the negative binomial models does not show significant results in the logistic regression model. One potential cause is that cluster masks the difference in drivers in the low risk group, which constitutes the majority of the drivers. We presented the full model to be consistent with the negative binomial model.

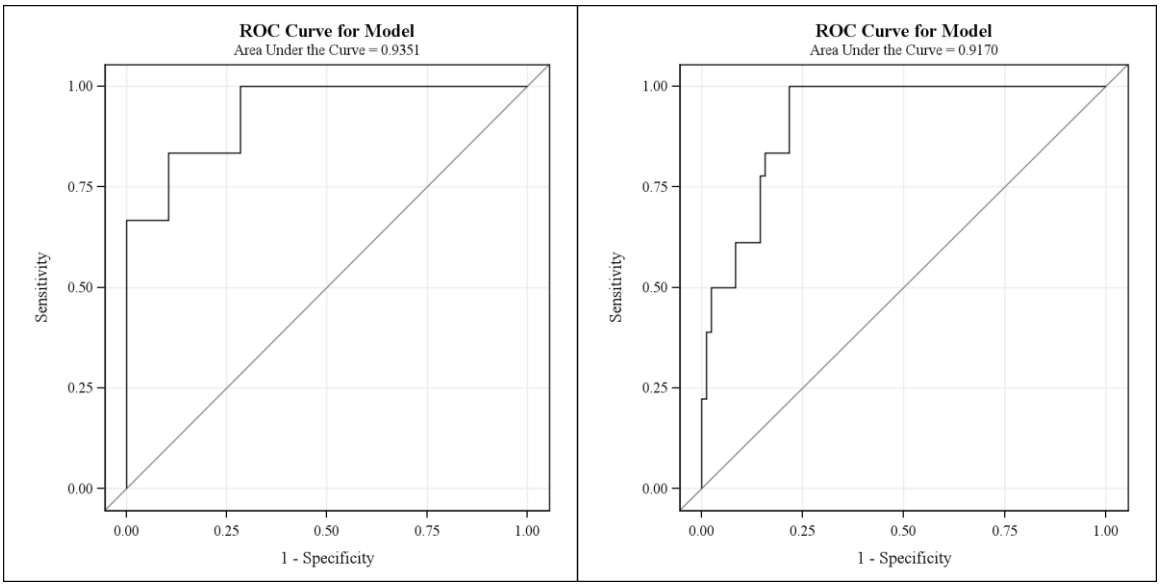
The predictive formulas are as follows,

$$\textbf{Model 1}: Probability(high\ risky\ driver) = \frac{\exp(-3.07 + 0.32 \times CIR - 5.86 \times Age25T55 - 2.0 \times AgeL25 - 0.033 * EXT)}{1 + \exp(-3.07 + 0.32 \times CIR - 5.86 \times Age25T55 - 2.0 \times AgeL25 - 0.033 * EXT)}$$

$$\textbf{Model 2}: Probability(high\ or\ moderate\ risky\ driver) = \frac{\exp(-1.92 + 0.27 \times CIR - 2.29 \times Age25T55 - 0.22 \times AgeL25 - 0.032 * EXT)}{1 + \exp(-1.92 + 0.27 \times CIR - 2.29 \times Age25T55 - 0.22 \times AgeL25 - 0.032 * EXT)}$$

where CIR is the CI rate, EXT is the extroversion factor, $Age25T55$ is a dummy variable equaling 1 if the driver age is between 25 and 55 and 0 otherwise, $AgeL25$ is a dummy variable equaling 1 if the driver age is less than 25 and 0 otherwise,

To evaluate the prediction performance, ROCs for both models were generated as shown in Figure 3. Both models showed very strong predictive power. The AUC for model 1 is 0.9351 and 0.917 for model 2, all close to the perfect AUC value 1.



a. Model 1 ROC: High vs. Moderate/Safe b. Model 2 ROC: High/Moderate vs. Safe

Figure 3 The ROC Curves

SUMMARY AND DISCUSSION

Safety risk varies substantially among drivers. To identify factors associated with risk at individual drivers can have a significant impact on driver education strategy and developing safety countermeasure. This paper attempted to connect driving risk with driver demographic information, personality, and driving characteristics. The focus is to examine whether critical incident, a measure of driving characteristic, can be used to assess driving risk and predict high risky drivers.

The negative binomial regression indicated that critical incidents had a significant impact on individual driving risk with a 10% increase in CNC rate for every one unit of increase in critical incidents. The driver age group 25-55 was the safest group whose relative crash rate was approximate 60% of that of the >55 age group and 50% of the <25 age group. The NEO personality factor extroversion also showed a significant negative correlation with the CNC rate.

The cluster analysis indicated that about 6% of drivers had a substantial higher risk (10 times higher than low risk groups) and about 12% of the driver showed moderate to high risk (4 times higher than low risk group). The logistic regression models had a very high prediction power. Measured by the receiver operation curve (ROC), the area under the curve for predicting the high-risk group and moderate to high risk group were 0.935 and 0.917. The critical incident

again had a statistically significant influence in the prediction. For every one unit of increase in critical incident rate, the relative probability of being a high-risky or moderate-to-high risky driver increased by approximate 40% to 30%.

The results of this study confirmed that aggressive driving characteristics, as represented by critical incident rate, have a strong relationship with individual driving risk. Furthermore, a logistic regression-based prediction model using critical incident rate can successfully identify high and moderate risk drivers. This result can have significant influence on the individual risk assessment. As the number of accidents for individual drivers is often limited, predicting high risk drivers by past accident history could be inefficient. The results of this study indicated that critical incidents, which has about 100 fold higher frequency comparing to crash and 10 folder higher frequency to near-crash, is a powerful predictor for high risk drivers. This particularly important for developing proactive safety counter measures for improve the safety of high risk drivers.

There are several possible extensions to this study. First more sophisticated statistical approach, such as principle analysis and factor analysis, can be used to include highly correlated covariate in the prediction model. Second, the driver distraction behavior, i.e., use cellphone, eating, texting etc. has not been evaluated. It will be of interest to how those factors will influence individual driving risk.

REFERENCES

Tarko, A., G. Davis, Ni. Saunier, T.Sayed, S. Washington (2009) Surrogate Measure of Safety: White Paper. Transportation Research Board ANB20(3) Subcommittee on Surrogate Measures of Safety

Arthur W., Graziano W.G. (1996) The Five-Factor Model, Conscientiousness, and Driving Accident Involvement. Journal of Personality 64:593-618

Costa, P. T., and McCrae, R. R. (1992). *Revised NEO Personality Inventory and NEO Five Factor Inventory: Professional Manual.* Psychological Assessment Resources. Lutz, Florida.

Dahlen, E. R. and R. P. White (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. and B. N. Fildes (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., Perez, M. A., Hankey, J., 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. (Interim Project Report for DTNH22-00-C-07007, Task Order 6; Report No. DOT HS 810 593). Washington, D.C.: National Highway Traffic Safety Administration.

Guo, F., S. G. Klauer, J. M. Hankey, T. A. Dingus (2010a), "Using Near-Crashes as a Crash Surrogate for Naturalistic Driving Studies" the Transportation Research Record: Journal of the Transportation Research Board. Vol 2147 pp 66-74.

Guo, F. J. M. Hankey (2009), "Modeling 100-Car Safety Events: A Case-Based Approach for Analyzing Naturalistic Driving Data" (2009) the National Surface Transportation Safety Center for Excellence,

Guo, F., X. Wang, and M.A. Abdel-Aty (2010b), "Modeling Signalized Intersection Safety with Corridor Level Spatial Correlations", *Accident Analysis and Prevention.* 42(1): 84-92.

Hauer, E., Ng, J., and Lovell, J. (1988), "Estimation of Safety at Signalized Intersections," *Transportation Research Record: Journal of the Transportation Research Board,* 48-61.

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

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

Klauer, S.G., F. Guo, J. Sudweeks, and T. A. Dingus (2010), An Analysis of Driver Inattention Using a Case-Crossover Approach On 100-Car Data, the National Highway Traffic Safety Administration, report number: DOT HS 811 334,.

Klauer, S.G., Dingus, T. A., Neale, V. L., Sudweeks, J.D., and 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). Washington, DC: National Highway Traffic Safety Administration.

Li, W., Carriquiry, A., Pawlovich, M., and Welch, T. (2008), "The Choice of Statistical Models in Road Safety Countermeasure Effectiveness Studies in Iowa," *Accident Analysis & Prevention*, 40, 1531-1542.

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

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. and K. S. Sankey (2008). "Relationships between young drivers' personality characteristics, risk perceptions, and driving behaviour." *Accident Analysis & Prevention* 40(2): 541-547

Maze, T. H., Agarwai, M., and 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.

Mitra S., Washington S. (2007) On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis & Prevention 39:459-468

Olson, R.L., Hanowski, R.J., Hickman, J.S., & Bocanegra, J. (2009). *Driver distraction in commercial vehicle operations* (Report no. FMCSA-RRR-09-042). Washington, DC:, Federal Motor Carrier Safety Administration.

Poch, M., and Mannering, F. (1996), "Negative Binomial Analysis of Intersection-Accident *Frequencies*," *Journal of Transportation Engineering*, 122, 105-113.

Rothman K.J., Greenland S., Lash T.L. (2008) Modern epidemiology. 3rd ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia.

Shaw L., Sichel H.S. (1971) Accident proneness: Research in the occurrence, causation, and prevention of road accidents Oxford, England: Pergamon.

Segovia-Gonzalez, M. M., Guerrero, F. M., and Herranz, P. (2009), "Explaining Functional Principal Component Analysis to Actuarial Science with an Example on Vehicle Insurance," *Insurance: Mathematics and Economics*, 45, 278-285.

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.

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

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

Venezian, E. (1981), "Good and Bad Drivers-a Markov Model of Accident Proneness," *Proceedings of the Casualty Actuarial Society Casualty Actuarial Society*, LXVIII, 65-85.

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.