

# **EVALUATION OF THE CHARACTERISTICS OF DRIVERS WITH MULTIPLE CRASHES**

**FINAL REPORT**



Prepared by

Susantha Chandraratna  
and  
Nikiforos Stamatiadis

Department of Civil Engineering  
University of Kentucky

For the  
Southeastern Transportation Center  
USDOT Transportation Center

JUNE 2004

## TABLE OF CONTENTS

TABLE OF CONTENTS.....	i
LIST OF FIGURES .....	ii
LIST OF TABLES.....	iii
ABSTRACT.....	iv
1.0 INTRODUCTION .....	1
2.0 BACKGROUND INFORMATION / LITERATURE REVIEW.....	3
2.1 Problems with Past Research.....	8
3.0 METHODOLOGY .....	11
3.1 Databases.....	11
3.2 Data Set.....	12
3.3 Variable Selection.....	14
3.4 Modelling.....	15
3.5 Model Validation.....	17
3.6 Selection of Cut-point.....	17
3.7 Receiver-Operating Characteristic Curves .....	18
4.0 STATISTICAL ANALYSES AND RESULTS .....	21
5.0 DISCUSSION AND CONCLUSIONS .....	27
REFERENCES .....	32

## LIST OF FIGURES

Figure 1: Area under a ROC curve .....	20
Figure 2: Receiver Operating Characteristic Curve for the best fit model .....	24

## LIST OF TABLES

Table 1: Time gap between the last two crashes, Kentucky crashes, 2002 .....	9
Table 2: Crash frequencies of crash-involved drivers in 2002 .....	13
Table 3: 2002 Kentucky at-fault drivers' previous crash history .....	13
Table 4: 2002 Kentucky not-at-fault drivers' previous crash history .....	14
Table 5: General form of the crosstabulation output for a given cut-point .....	17
Table 6: Properties of diagnostic test.....	18
Table 7: Result of logistic regression.....	22
Table 8: Logistic regression model summaries .....	23
Table 9: ROC trade offs for the best fit model for the experienced drivers .....	24

## ABSTRACT

A goal for any licensing agency is the ability to identify high-risk drivers. Kentucky data shows that a significant number of drivers are repeatedly involved in crashes. The objective of this study is the development of a crash prediction model that can be used to estimate the likelihood of a driver being responsible for a near future crash occurrence. Multiple logistic regression techniques were employed using the available data for the Kentucky licensed drivers. This study considers as crash predictors the driver's total number of previous crashes, citations accumulated, the time gap between the latest two crashes, crash type, and demographic factors. The driver's total number of previous crashes was further disaggregated into the drivers' total number of previous at-fault and not-at-fault crashes. The overall efficiency of the model is 64.15 percent and the model can be used to correctly classify at-fault drivers up to 74.56 percent with an overall efficiency of 63.34 percent. The total number of previous at-fault crash involvements, and having previous driver license suspensions and traffic school referrals are strongly associated with a driver being responsible for a subsequent crash. In addition, a driver's risk is increased by: being very young or very old; being male; having accumulated both speeding and non-speeding citations; and having a recent crash involvement. Thus, the model presented here enables agencies to identify potential at-risk drivers. Moreover, the model can be used for driver control programs aimed at preventing road crashes such as from a simple form of issuing warning letters to a license suspension.

## 1.0 INTRODUCTION

A knowledge regarding the factors that increase risk for motor vehicle crashes is important to improve traffic safety. This is more important when the social and economical impacts of vehicle crashes are taken into consideration. The total number of traffic collisions is gradually increasing on Kentucky roads (1). During the 1999 to 2002 period, the annual total number of police reported crashes in Kentucky increased from 132,216 to 153,921. In 2002, 810 of these crashes were fatal. Estimates also show that the 2002 Kentucky crashes resulted in the loss of \$5.9 billion worth of quality of life associated with deaths and injuries as well as the economic costs (1). The national figures are more alarming. In 2001, Federal Highway Administration (FHWA) claimed that motor vehicle crashes are the leading cause of death in the U.S. for people between the ages of 6 and 33 (2). Vehicle crashes are a greater threat to human life and health in the U.S. than that of crime. In 1999, there was one murder every 34 minutes and one violent crime every 22 seconds, whereas in the same year there was one fatality in a vehicle crash every 13 minutes and one injury-crash every 15 seconds. Annually, these crashes account for more than 1 million years of potential life lost before the age of 65. Each year more than 40,000 people are killed and more than 5 million are injured on U.S. highways. It is estimated that the cost of vehicle crashes in the United States is to be \$150 billion a year, or 2.2 percent of the Gross Domestic Product (2).

Whenever one drives a motor vehicle, there is always some degree of risk of being involved in a crash. In some crashes, a vehicle problem or an environmental condition may significantly contribute to a crash and then such crashes may be considered as a result of chance. However, this is not the case for drivers who had been repeatedly involved in crashes. As early as 1920, Greenwood and Yule hypothesized that some people have many more crashes than can be expected by chance, so these people are considered as crash-prone (3). After analyzing a Spanish bus drivers' crash database during 1976-1983, Blasco et al. concluded that drivers' recurrent crashes primarily occurred because of human error rather than by chance (4). Therefore, drivers who have been involved in multiple at-fault crashes could be generally considered as high-risk drivers.

To preserve life and prevent injury and economic costs, federal and state agencies are implementing many driver control programs aimed at preventing road crashes. These programs vary from a simple form of issuing warning letters to license suspension (i.e. the driver license is temporarily withdrawn) or revocation (i.e. the driver license is terminated). In addition to these two extremes, many states also use methods like educational brochures, group educational meetings, diagnostic reexaminations, individual counseling, and administrative hearings aimed at improving traffic safety. Primarily, the main subjects of such programs are drivers who are considered as habitual traffic violators and crash repeaters. These driver programs not only incur substantial costs to agencies but also have the potential to affect an individual's freedom to travel. Thus, it is of high importance to identify the factors that increase risk for motor vehicle crashes and then to identify the drivers targeted for these programs correctly and accurately.

Many researchers have attempted to determine which variables are the best crash predictors and revealed that a number of demographic (e.g., age, sex), psychological (e.g., aggression), situational (e.g., city size), and behavioral (e.g., risky driving practices, road violations) factors increase an individual's risk for crashes for the driving population in general (5,6,7). Most of the literature indicates that drivers who have driving records with citations, crashes, or both could be generally considered as crash-prone drivers. A substantial body of research has specifically focused on identifying crash-prone drivers using their past crash and citation records. However, most of these studies were only aimed at identifying crash-prone drivers irrespective of their responsibility for the particular crash. Being involved in a crash without any responsibility can be considered a random event (8,9). Thus, this study considers only drivers who are responsible for crashes as high-risk drivers. With this in mind, this research effort focuses on predicting the likelihood that a driver will be responsible for a recurrent crash occurrence in the future using driver records in the 1995-2002 period. Multiple logistic regression techniques were employed for the subjects who were repeatedly involved in crashes in order to establish relationships between driver responsibility and potential crash predictors. It should be noted from the outset that this could be considered as a conditional probability, i.e. determining the culpability given the fact that a crash has occurred. However, the ability to identify at-fault drivers in advance is useful because they are the ones who are responsible for originating crashes. Thus, the model presented here enables licensing agencies to develop possible remedies and alert risky drivers of their potential to be involved in a crash.

## **2.0 BACKGROUND INFORMATION / LITERATURE REVIEW**

Over the past decade, significant efforts have been completed that aimed to assist licensing agencies and improve traffic safety. Most of the past studies aiming to identify high-risk drivers had used either their past driving records or driver behavior questionnaires or both. The analysis of driver history and records is the most popular approach. The probability of drivers' involvement in recurrent crashes merely by chance is minimal and a significant fraction of traffic violations and crash risk is attributable to factors under an individuals' control. Therefore, drivers who were involved in several crashes are considered high-risk drivers (4,10). More accurate driver records can be obtained from the relevant agencies than data relying on questionnaire. Past studies have also claimed that other personal and impersonal factors and variables contributed to a driver being involved in a crash (11). Thus, an extensive literature review, which focuses on the studies aimed to estimate an individual driver's future crash potential by using drivers' prior driving record histories, was undertaken to identify those factors and variables.

The California Driver Study issued by the California Department of Motor Vehicles in 1958 is considered to be the first such study (5). Thereafter, a substantial body of research had been focused on establishing detailed estimations of a driver's future crash potential on the basis of prior driving record histories (5,6,10). Most of these studies found a statistically significant relationship between the number of crash involvements and the number of traffic convictions. However, after a much more extensive study initiated by CDMV in 1964, Peck et al. concluded that the statistical nature of driver crash frequencies make it impossible to accurately predict which individuals will and will not be involved in crashes (10). They pointed out that they were not able to predict more than 15 percent of the variance in the three year crash frequencies of the California male driving population. The efficiency of their prediction model obtained for females were even lower than that of the model for males. This may be due to the possible significant lower exposure of females as compared to males in the early 60's. However, Peck and Kuan continued the study by limiting it to a subset of drivers who had been licensed for at least six years (5). The objectives of this subsequent study were: (1) to determine the relative importance of territory, prior driving record and other variables in predicting future crash involvement; and

(2) to determine whether a driver's area of residence is a fair and actually sound rating factor. Using two separate random samples totaling more than 90,000 drivers, various prediction models were developed using multiple regression techniques in order to predict subsequent three year crash involvement frequency. They classified driver records by using multiple regression models with respect to different prediction cut-points ranging from 0.120 to 0.48. Although both territory and prior driving record proved to have some validity in predicting a driver's crash risk, the accuracy of the prediction was low with multiple correlations ranging from 0.08 to 0.25. Prior driving record, particularly a driver's previous number of traffic convictions, was a much better predictor than a territory, which may be less important than initially considered.

Lui and Marchbanks determined a relationship between previous traffic infractions and fatal automobile crashes by studying the Fatal Accident Reporting System (FARS) from 1984 to 1986 period (12). They suggested that the involvement in a fatal automobile crash is not a random event. In this study, they attempted to identify the time that elapses between the fatal crash and either the date of the last crash, suspension or conviction. Their results showed that the mean time between a previous traffic infraction and a fatal automobile crash was the shortest for individuals aged 16 to 25 years, who had a mean recurrent time of 14.2 months with a 90 percent confidence interval of 13.9 months to 14.6 months. Approximately 97 percent of the recurrent times occurred within 60 months of a given traffic infraction, with the highest risk of a fatal crash from three to seven months following the infraction.

Hauer et al. examined several tools for the identification of drivers who are most likely to have a crash in the near future and investigated how they perform using a four-year record for a large sample of Ontario drivers (13). They defined the individual's crash potential as the expected number of crashes per unit time and compared a variety of 16 different models including: (1) models that used age and gender as variables and those that did not; (2) models in which the number of at-fault, not-at-fault, or total crashes were defined as variables and models that make no use of this information; and (3) models that had a separate parameter for all convictions. They suggested that a weight of one to a conviction and 1.88 to a crash can be used in estimating the expected number of crashes per unit time without much loss in estimation accuracy. It was found that the model that made use of detailed information on age, gender, number of each of 14 types

of convictions, and number of at-fault crashes and not-at-fault crashes, was the most efficient model in terms of explaining the variance of estimated crash potential. The authors concluded that: (1) to identify drivers with crash potential, using only demerit points based on the perceived seriousness of convictions is not sufficient; (2) it is important to make use of the driver's previous crash record; (3) models that use only at-fault crashes perform worse than those that use all crashes; and (4) models that use age and gender perform better than the corresponding models that make no use of age and gender variables.

Chen, Cooper, and Pinili employed logistic regression analysis to identify drivers who were most likely to have one or more at-fault crash involvements after their crashes based on their records prior to their at-fault crash involvements (14). After analyzing 1,998,347 British Columbian drivers' records, they found a consistent increase in postperiod crashes per driver with increasing preperiod numbers of both crashes and convictions. The model they developed had a 48.7 percent efficiency in identifying the top 1,000 high risk drivers who had the highest probability of being culpably involved in crashes in the following two years. Furthermore, they indicated that a model that makes use of prior at-fault crash information can identify up to 23 percent more drivers who will have one or more at-fault crash involvements in the next two years than a model that uses conviction alone.

Analyzing Kentucky crash data during 1993-1997, Stamatiadis et al. indicated that there were 14,750 drivers per year with one conviction and one crash within one year, 20,300 drivers with one conviction and one crash within two years, 1,400 drivers with a conviction and two crashes within a year, and 3,100 drivers with a conviction and two crashes within two years (15). Furthermore, their research suggested that these high frequencies reveal a significant relationship between crashes and convictions. More interestingly, they pointed out that two percent of the drivers per year renewing their license would match the risk category of two convictions and a crash within two years. Although this percentage may not seem significant, for the long run these findings could be used in establishing criteria for requiring drivers to undergo additional tests prior to renewing their license.

At the same time as the Kentucky study, Gebers studied California's negligent-operator point system in which each conviction of a violation of the traffic law carries a specific number of negligent-operator points (6). He extracted a one percent random sample of the California driving population in 1992, and for each subject, collected demographic factors age and gender and driver record information, such as total crashes, total citations, at-fault crashes, negligent-operator points, and individual violation types. After studying 17 logistic regression models, he concluded that all of the models were consistent in demonstrating that increased probability of subsequent crash involvement is associated with increased prior citation and prior crash frequencies, being young, and being male. The model that used age, gender, license class, total number of citations, total number of crashes, and total number of failure-to-appear in court violations yielded 25,884 total crash hits during the next four years. Similarly, the hits for the model that used age, gender, license class, one parameter each for the number of zero to two-point citations were 25,881. Moreover, the correctly classified percentages for the crash-involved drivers by these two models were 27.31 percent and 27.57 percent, respectively. However, Gebers also noted that models that use only prior at-fault crashes as a predictor do not perform as well as models that use prior total crashes as a predictor. By deploying canonical correlation techniques in a subsequent research effort, Gebers and Peck achieved an accuracy level of 27.2 percent from their best model in identifying crash-prone drivers as well (16).

Some research suggested that drivers' unsafe acts should be sub-divided into three classes of behaviors: lapses (e.g. getting into the wrong lane when approaching a roundabout or a junction,) errors (e.g. fail to notice that pedestrians are crossing when turning onto a side street from a main road) and violations (e.g. speeding 26 or more mph over speed limit) (11,17). Mesken et al. pointed out that violations are more often reported by men than women while errors are more often reported by women than men (11). However, the men's violations declined with age but the women's errors did not. In their study, they further differentiated violation in to two groups, interpersonal violations and speeding violations, and revealed that these four types of conviction groups -- lapses, errors, interpersonal violations and speeding violations -- are statistically significant predictors of crashes.

In addition to driver history, age and gender have long been identified as crash predictors by

most of previous studies aimed to determine which variables are the best crash predictors. Studies repeatedly pointed out that young drivers are significantly over-represented in road crashes. The per-mile crash rate of 16-year-old drivers is approximately 10 times that of adults' crashes in which young drivers are involved (18). Examining the Kentucky crash database, Kirk and Stamatiadis revealed that crashes at intersections, rear-end crashes, crashes resulting from passing maneuvers, and single vehicle crashes are the most prominent crash types for young drivers (19). However, the results showed that for all crashes there is a general trend of decreasing involvement with increasing age, which indicates that their inexperience is the largest single contributor to their increased crash rates. Similar to young drivers, older drivers also show over-representations in road crashes. Kim et al. estimated that very young and very old drivers face up to three times the risk of being at fault compared to middle-aged drivers (20). Furthermore, Stamatiadis and Deacon have shown that elderly drivers are also more frequently involved in specific types of crashes (8). Another research, which also used the Kentucky crash database, revealed that older drivers (age>65 years) are more likely to be involved in left turn crashes compared to other drivers and the risk of their involvement in left turn crashes increases 1.08 times each year the driver ages (21). Elderly experience difficulties in maneuvers related to gap acceptance for crossing non-limited access highways, and high speed lane changes on limited-access highways compared to young drivers.

Similar to driver age, the relationship between drivers' gender and crashes has also been studied extensively. Massie et al. revealed that men do have a consistently higher risk of crash involvement per mile driven than women for the six combinations of crash severity and light condition they examined (22). However, it should be noted that Forward et al. compared the results of studies between 1970 to 1984 with those of between 1985 and 1997 on crashes and driving habits of males and females and concluded that the gender differences had decreased over time (23).

In addition to age and gender, variables like vehicle type and age, crash location, time of the day, weather condition, highway type etc. are proven to be reasonable crash predictors in various studies (24). Understandably, predicting driving risk by using age, gender and other impersonal

variables like weather condition would not allow an estimation of an individual driver's future crash potential rather than that of a group of drivers in question.

## **2.1 Problems with Past Research**

This literature review found no study that considered the drivers' total number of not-at-fault involvements as a predictor variable in identifying risky drivers. For example, Gebers attempted to identify crash-prone drivers by considering the drivers' total number of previous crashes and at-fault crashes separately as future crash predictors, but ignored the total number of previous not-at-fault crashes along with the total number of previous at-fault crashes (6). This procedure created some methodological problem for the models because they are determined in predicting not only future at-fault involvement but also not-at-fault crash involvement either by using only the total number of previous crashes or by using only the at-fault crashes as an independent variable.

All studies examined here were performed by subdividing the crash period analyses into two intervals in order to make a prediction of subsequent crash risk during a predetermined period based on data collected over a fixed *pre* period. For example, Chen et al. used the first three-year (i.e. 1985-1988) complete record of at-fault crash involvements and conviction records to identify the high-risk drivers who are most likely to be culpably involved in crashes in the following two years (i.e. 1989-1990) (14). Then, they carried out statistical procedures to develop models in order to determine the relationship between prior convictions and/or prior crashes and subsequent crashes. However, the Kentucky crash database shows that over 45 percent of drivers who had only two crash involvements had their second crash less than two years later, and 28.37 percent drivers had longer than a four-year time gap in between two crashes (Table 1). In addition, a significant number of consecutive crashes would not be identified for the drivers who had more than two crash involvements if the crash database was simply subdivided into two periods. Thus, the subdivision of the crash period into two intervals would not allow for properly identifying a driver's previous crashes and citations, since a significant number of subsequent crashes could have occurred only within either interval. Table 1 shows that up to 45 percent of the drivers' most recent history may be ignored in the analysis

as a result of such subdivisions. Thus, this study attempts to predict the driver's last involved crash by using his/her previous records. It is assumed that this procedure will improve the predicting power of the model because it uses more of the latest available data than the other procedures. For example, if a driver had two crashes in 2002, this procedure allows the use of the driver's first crash information in predicting the second crash within the same year.

<b>Time gap (years)</b>	<b>Frequency</b>	<b>Percentage</b>	<b>Cumulative percentage</b>
1 or less	17,744	25.66	25.66
1-2	13,436	19.43	45.09
2-3	10,377	15.01	60.09
3-4	7,979	11.54	71.63
4-5	6,532	9.45	81.08
5 or more	13,087	18.92	100.00

**Table 1: Time gap between the last two crashes, Kentucky crashes, 2002**

Past studies assumed that a driver's probability of being responsible for a future crash anytime during a two or three-year period is the same. However, a driver's ability may change with time or drivers may change their behavior. For example, young drivers may improve their driving skills over a three year period more than they might in just six months after an at-fault crash incident, or drivers may drive more defensively immediately after being involved in a crash. Moreover, it could be hypothesized that a driver who was involved in a crash is likely to drive less in the following few years. This may be due to reasons such as injuries, bad health conditions, psychological impact, having no vehicle to drive, or even a driver license suspension because of the crash. Thus, aforesaid study procedure allows for evaluating the effect of the time gap between the current crash in question and the penultimate crash of the same driver.



## 3.0 METHODOLOGY

### 3.1 Databases

Two Kentucky databases were used in this analysis for the 1995-2002 period. The Kentucky Driver License (KyDL) database for the year 2002, which contains 3,201,620 driver records, was used to extract driver license number, age, gender and type of citation up to 1998. However, the KyDL database does not provide a detailed crash history. Thus, license number, crash date, crash type and human factors for each respective crash were extracted from the Kentucky Crash (KyC) database, which consists of police reported Kentucky crashes in the 1995-2002 period.

Citations remain in the KyDL database for only a five-year period and they are then purged. Therefore, driver records for the 1995-1997 period were extracted from the 1999 records of KyDL database and merged with the 2002 records of the KyDL database. With relation to crashes, not all citations in the KyDL database may be indicative of *risky behavior* for the reason that these convictions include offenses like “*no liability insurance in force.*” Therefore, such citations, along with all other non-moving violations, were considered *no-risk citations* (NORISK) for this study. All other moving violations were considered as *risky* behavioral violations. It should be noted that these risky violations are reported under 83 categories in the KyDL database. However, Mesken et al. pointed out that these risky behavioral citations can be categorized into four major groups: lapses (LAPSES), errors (ERRORS), non-speeding violations (VIOLATE) and speeding violations (SPEEDING) (11). Lapses can be defined as “the unwitting deviation of action from intention” (e.g. failure to illuminate headlights), while errors can be defined as “the failure of planned actions to achieve their intended consequences” (e.g. improper turn) (25). Thus, some researchers consider lapses as inattention errors and inexperience errors (11). On the other hand, both non-speeding and speeding violations can be defined as “the deliberate deviations from those practices believed necessary to maintain the safe operation of a potentially hazardous system” (e.g. disregard of stop sign, all types of speed violations) (25). This method was adopted here to group 83 citation types into these four categories. It should be noted that even though Kentucky has a “Traffic School” system, where citations are eliminated because of Traffic School attendance, the citation information is

provided in the database but with no points. These citations were included here because the study focused on each driver's total number of citations instead of accumulated points. For Kentucky, the accumulation of 12 points within a two-year period results in the suspension of the driver's license. License suspension can also occur if a driver is cited for violation of 26 mph over speed limit on any road/highway, attempting to elude police officers or racing. However, a significant number of drivers who have had their license suspended were involved in crashes after having their driver's license reinstated. The same was observed for the drivers who were referred to traffic school. Thus, the effects of traffic school attendance (SCHOOL) and driver license suspension (SUSPENSE) were also tested here.

The KyDL database maintains driver records under the driver license number whereas the KyC database maintains crash records under a *master file number* that makes available crash information for all the drivers involved in crashes. However, the unit of analysis in this study is the driver involved in a crash. Therefore, if a crash involved two or more vehicles, the information for each driver was disaggregated and separate records for each driver were created.

### **3.2 Data Set**

The two databases were merged by matching the driver license number. The merged data shows that there were 157,832 Kentucky licensed drivers involved in reported crashes within the state in 2002. Table 2 shows the crash frequencies of these drivers during the 1995-2002 period. Since the objective of this research is limited to predicting the likelihood of a driver being responsible for a recurrent crash involvement, only drivers who had at least two crash involvements are selected as a sample for the statistical analysis, producing a data set of 69,155 drivers (i.e. 43.82 percent of the 2002 Kentucky crash-involved driver population). Limiting the selection to 2002 Kentucky crash-involved driver ensures that each crash driver would have an eight-year citation record and crash record.

Total number of crash involvements <sup>a</sup>	Number of drivers	Percentage
1	88,677	56.18
2	43,525	27.58
3	16,499	10.45
4	4,473	2.83
5 or more	4,658	2.95

<sup>a</sup> the last crash is also accounted for the total number of crash involvements.

**Table 2: Crash frequencies of crash-involved drivers in 2002**

Kentucky crash investigating officers identify reasons or factors that could have potentially contributed to the crash occurrence (e.g. unsafe speed, failed to yield right of way, alcohol involvement, etc) and assign them to the drivers involved. These indicators can be extracted from the Kentucky crash database under the human factor category and were used here to determine responsibility. Using this information, most of the drivers involved in Kentucky crashes can be categorized as either at-fault drivers or not-at-fault drivers. This approach has been tested previously and does not introduce any bias in the analysis (8,9). At-fault drivers are defined as those drivers who were cited as having one or more human factors contributing to the crash. On the other hand, not-at-fault drivers are defined as drivers who were not cited as having human factors contributing to the crash. Missing data of 1,154 drivers were observed for human factors in the KyC database.

Out of the 34,932 (Table 3) drivers who were at-fault in 2002, 22,518 (64.46 percent) had at least one or more previous at-fault crashes with different combinations of not-at-fault crashes. In addition, 16,842 (48.21 percent) of them had a higher number of previous at-fault crashes than not-at-fault crashes, while 13,828 (39.59 percent) had a higher number of previous not-at-fault crashes. Similar trends were observed for the not-at-fault drivers in 2002 (Table 4).

		Total Number of previous not-at-fault crashes					Total
		0	1	2	3	4 or more	
Total Number of previous at-fault crashes	0	-	10,095	1,902	323	94	12,414
	1	11,669	3,882	996	215	62	16,824
	2	2,601	1,102	344	92	25	4,164
	3	595	333	129	34	22	1,113
	4 or more	182	127	57	30	21	416
Total		15,047	15,539	3,428	694	223	34,932

**Table 3: 2002 Kentucky at-fault drivers' previous crash history**

		<b>Total Number of previous not-at-fault crashes</b>					
		<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4 or more</b>	<b>Total</b>
<b>Total Number of previous at-fault crashes</b>	<b>0</b>	-	11,72	2,283	419	93	14,523
	<b>1</b>	9,735	3,631	894	229	77	14,566
	<b>2</b>	1,820	869	259	73	40	3,061
	<b>3</b>	341	212	89	29	19	690
	<b>4 or more</b>	77	73	46	18	15	229
	<b>Total</b>	11,973	16,513	3,571	768	244	33,069

**Table 4: 2002 Kentucky not-at-fault drivers' previous crash history**

### 3.3 Variable Selection

It can be hypothesized that the probability of any driver being involved in a crash without having any responsibility is the same because an at-fault driver does not intentionally select whom to strike. Thus, this study assumes that these not-at-fault drivers represent the general driver population. However, a number of California studies do not agree with this assumption (6,26). They argue that these not-at-fault involvements are not random events. This may be because drivers who are less capable may be more likely to be involved in crashes due to their less effective defensive driving techniques, even though they are not responsible for the initial cause of the crash. However, it is true that the drivers who are more exposed to traffic hazards, have a higher chance of being involved in a crash without being responsible. Therefore, it is of interest to examine whether high-risk drivers, with different combinations of previous at-fault and not-at-fault involvements, are responsible for their subsequent crash involvements. Thus, this study uses not-at-fault driver data to develop a model that could predict the driver's risk of being at-fault against being not-at-fault for a crash occurrence.

In determining the model, the total numbers of previous at-fault crashes (ATFAULT) and not-at-fault crashes (NOTFAULT) of the 69,155 drivers along with citation data -- NORISK, LAPSES, ERRORS, VIOLATE, SPEEDING, SCHOOL, SUSPENSE -- were used as the primary crash predictors. NORISK, LAPSES, ERRORS, VIOLATE, and SPEEDING were considered as continuous variables. On the other hand, suspensions and traffic school referrals were considered as categorical variables and were coded as one (1) for presence and zero (0) otherwise.

However, it should be noted that citations accumulated after the last crash in the sample were excluded, since those citations cannot be used as predictors for the crash in question. Similarly, convictions accumulated because of crashes were also disregarded to avoid duplicating the crash data.

Age (AGE) and gender of the driver (GENDER) have also been used in the past as good crash predictors, and they were examined here as well (6, 10, 12, 16). While different factors may come into play over the time gap between two consecutive crashes, this study hypothesizes that high-risk drivers may be more prone to have frequent crashes within shorter time intervals. Thus, the time gap (TIME\_GAP) between the last two crashes is also tested as a crash predictor.

Crash type (CR\_TYPE) presumably also plays a role in crashes. Sometimes crashes may be avoided by not-at-fault drivers if they act defensively. In terms of the complexity, intersection maneuvers may be the most difficult for any driver in general. Thus, this study also examines how high-risk drivers may be responsible in intersection crashes (TYPE\_1) compared to non-intersection related crashes (TYPE\_2). Drivers are more likely to be victims of circumstances beyond their controls such as when they collide with an animal (TYPE\_3). On the other hand, drivers in the crash types like “vehicle overturning,” “collision with fixed objects” and “off road crashes” may not be able to blame other drivers on the road (TYPE\_4).

### **3.4 Modelling**

Logistic regression has been proven the most appropriate statistical technique for this type of modelling (6, 8, 12, 27). The reason is that logistic regression is a form of regression, which is used when the dependent variable is a binary; that is it can have only two values (at-fault and not-at-fault) (28,29,30). Logistic regression technique is particularly advantageous when the effects of more than one independent variable are important. These independent variables can be continuous variables, categorical variables or both. In addition, logistic regression does not assume linearity of the relationship between the independent variables and the dependent variable, does not require normally distributed variables, and in general has less stringent

requirements than linear regression. In this analysis, the dependent variable is the fault status of the driver. Thus, the probability of occurrence of a crash is modeled as follows:

$$\text{Prob}(at - fault\ driver) = \frac{1}{1 + e^{-z}} \quad (\text{Eq.1})$$

Where Z is the linear combination

$$Z = B_0 + B_1X_1 + B_2X_2 + \dots + B_NX_N$$

and the B's are coefficients, estimated using the maximum-likelihood method, and the X's are the independent predictor variables previously discussed. All variables were treated as continuous variables except SCHOOL, SUSPENSE, GENDER and CR\_TYPE which were treated as categorical variables.

For the logistic model, Eq. 1 is commonly rearranged as shown in Eq. 1A to easily understand the interpretation of the logistic coefficients.

$$\log\left(\frac{\text{prob}(at - fault\ driver)}{\text{prob}(not - at - fault\ driver)}\right) = B_0 + B_1X_1 + B_2X_2 + \dots + B_NX_N \quad (\text{Eq.1A})$$

The null hypothesis is that all the coefficients in the equation take the value zero. The null hypothesis can be rejected in the statistical sense if the relevant model parameter(s) were statistically different from zero at a level of significance of 0.05. The SPSS statistical software was used throughout this study.

An advantage of the models derived by the logistic regression, besides the ability of predicting the probabilities of drivers being at-fault, is that with all other predictor variables held constant, the risk increase for every one unit increase in each predictor variable can be estimated. This increase is known as log odds and is equal to the corresponding  $B_i$  coefficient. Since odds ratios are more useful to interpret models rather than log odds, the logistic equation can be written in terms of odds ratios (Eq. 1B).

$$\frac{\text{prob}(at - fault\ driver)}{\text{prob}(not - at - fault\ driver)} = e^{B_0 + B_1X_1 + \dots + B_pX_p} \quad (\text{Eq. 1B})$$

### 3.5 Model Validation

Since the objective of this study is to develop a model to identify risky drivers, the model needs to be able to correctly classify the at-fault drivers from a new data set. Thus, attempts must be made to validate the model with data which were not used in its development. Since Kentucky has large amounts of driver data, this can be simply achieved by the *Holdout Data Procedure* which has been proven to produce unbiased estimates of the probabilities of correct classifications (28). A holdout data set is one that can be used to validate a model but is not used to develop the model. For this step, two data sets were developed where each driver was randomly allocated to one of the two sets. Finally, one set, which is known as *calibration data*, was used for the model development and the other holdout dataset for the model validation. Although one drawback of this method is that all the data are not being used to develop the model, any possible adverse effects can be considered negligible when the calibration data set is so large (28).

### 3.6 Selection of Cut-point.

The logistic regression can only be used to estimate the probability of occurrence of a crash with respect to input values for the independent variables. Thus, a cut-point is necessary to assign drivers with values less than or equal to the cut-point as not-at-fault and drivers with values greater than the cut-point as at-fault drivers. Thus, it is understood that the number of drivers predicted as at-fault or not depends on the cut-point selected by the analysis. Since the usefulness of a model depends on how many more at-fault drivers it can accurately classify than would be expected by chance alone, sensitivity analysis was performed to select the optimal cut-point for the model. The general form of crosstabulation of predicted and actual outcome for a given cut-point is shown in Table 5.

		Predicted outcome	
		At-fault	Not-at-fault
Actual outcome	At-fault	A (true positive)	B (false negative)
	Not-at-fault	C (false positive)	D (true negative)

**Table 5: General form of the crosstabulation output for a given cut-point**

If the developed model is perfect, zero values should be observed for B and C in the crosstabulation. However, a perfect prediction for crash involvements is impossible because human related events are not bound by specific factors. Therefore, the best model is considered here as the one which produces reasonably low error values for B and C while identifying as many risky drivers as possible. To this end, the commonly accepted procedures in the sensitivity analysis were adopted and the definitions and relevant formulae are given in Table 6.

<b>Term</b>	<b>Description</b>	<b>Formula</b>
Sensitivity (True positive rate)	The proportion of the crash-involved drivers that were correctly classified to be crash-involved by the model.	$Sensitivity = A / (A+B)$
Specificity (True negative rate)	The proportion of crash-free drivers that were correctly classified to be crash-free by the model.	$Specificity = D / (C+D)$
Efficiency of the model	The proportion of both crash-involved and crash-free drivers correctly classified by the model.	$Efficiency = (A+D) / (A+B+C+D)$
False negative rate	The proportion of crash-involved drivers who are erroneously classified to be crash-free.	$False\ negative\ rate = 1 - sensitivity$
False positive rate	The proportion of crash-free drivers who are erroneously classified to be crash-involved.	$False\ positive\ rate = 1 - specificity$

**Table 6: Properties of diagnostic test**

Although the goal is to minimize error values for the false negatives and false positives, these two erroneous classifications are unfortunately, reciprocally related. Any decrease of the false negative rate causes an increase in the false positive rate. Thus, the optimal cut-point for identifying at-fault drivers has to be established by compromising between these two errors.

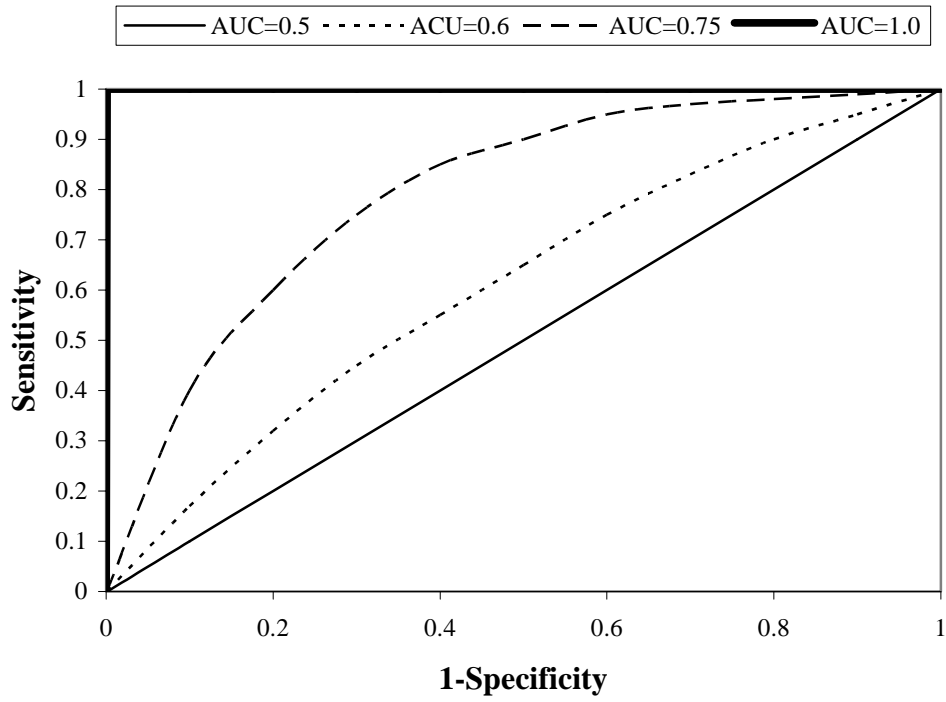
### **3.7 Receiver-Operating Characteristic Curves**

Receiver-Operating characteristic Curves (ROC Curves) is one technique which has been widely used for diagnostic accuracy of models since the early 1950s (31). The diagnostic accuracy means the quality of the classification between at-fault and not-at-fault drivers by the model in question. ROC curves provide a pure index of accuracy by demonstrating the limits of the model's ability over the complete range of cut-points. This is achieved by plotting all the

possible values of sensitivity versus the respective values of 1- specificity for the complete range of decision thresholds of a given study's results.

Theoretically, the area under ROC curves (AUC) varies from 0.5 (when no apparent distributional difference between the two driver groups can be achieved by the model's predictions) to 1.0 (when perfect classification of two driver groups is possible). This means that a ROC curve lies on the main diagonal when the model prediction power is at the level of chance and its true and false rates are equal across the range of possible cut-point values. As the model prediction power increases, the distance of the observed ROC curve from the chance line increases. When the ROC curve passes from the lower left corner through the upper left corner to the upper right corner, the AUC reaches its maximum of 1.0. This can be graphically shown as in Figure 1 for several ROC curves with their corresponding AUC values. A value less than 0.5 may be experienced because of an incorrect decision rule for the driver classification and can be corrected by reversing the rule.

Since the AUC does not depend on a particular point (e.g. a desired cut-point) or a range, the AUC is considered to be the most common global measure to quantify the diagnostic accuracy of a model (32). For example, an area of 0.7 means, by rewording Zweig and Campbell's medical example, that a randomly selected individual from the at-fault driver group has a predicted probability larger than that for a randomly chosen individual from the not-at-fault driver group 70% of the time (32). However, it does not mean that an identification of risky driver occurs with a probability of 0.70 nor that an identification of at-fault driver is associated with an actual occurrence 70% of the time.



**Figure 1: Area under a ROC curve**

#### 4.0 STATISTICAL ANALYSES AND RESULTS

A calibration data set of 34,684 drivers was randomly generated from the 69,155 available driver data for this analysis. This allows validating the model with a holdout dataset of 34,471 drivers. Then, for the Kentucky drivers' last crashes in 2002, driver responsibilities (i.e. at-fault or not) were used as the dependent variable and logistic regression was adopted to determine the best fit model between the dependent variable and the independent variables. All the coefficients of the model were tested based on the Wald Statistic. If a variable was not significant at 0.05 level, then the variable was removed from the model and the test was run again. It was observed that ten independent variables -- ATFAULT, NOTFAULT, SCHOOL, SUSPENSE, VIOLATE, TIME\_GAP, AGE, GENDER and CR\_TYPE --were significant at 0.05 level for this model whereas the variables NORISK ( $p=0.070$ ), LAPSES ( $p=0.881$ ) and ERRORS ( $p=0.837$ ) were not. The summary of the model determined after removing all non-significant variables is given in Table 7.

The model was then manually tested for different combinations of significant variables rather than relying on the SPSS available methods (e.g., forward stepwise selection and backward elimination) to determine whether the model efficiency can be improved. If TOFAULT was removed from the model, the log-likelihood increased from 40,372.94 to 40,467.26 which is significant at 0.05 level. This means the removal of TOFAULT adversely affects the model's efficiency, and thus TOFAULT should not be excluded from the model. The log-likelihood value, the Cox-Snell  $R^2$ , the Nagelkerke  $R^2$ , and AUC value obtained by the removal of each variable are shown in Table 8. In addition, the values of the model with all the significant predictor variables are also included in the same table for the comparisons. It can be seen that the removal of any variable causes the log-likelihood value to increase significantly, and thus the exclusion of any variable from the model reduces the possibility to better describe the effects of predictor variables on identifying future at-fault drivers.

Predictor variable	Regression coefficient	Standard error	Wald Statistic	p-value	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
ATFAULT	0.1536	0.0159	92.80	0.000	1.17	1.13	1.20
NOTFAULT	-0.0510	0.0162	9.90	0.002	0.95	0.92	0.98
SCHOOL	-0.1300	0.0283	21.15	0.000	0.88	0.83	0.93
SUSPENSE	-0.1525	0.0439	12.06	0.001	0.86	0.79	0.94
VIOLATE	0.0675	0.0218	9.59	0.002	1.07	1.03	1.12
SPEEDING	0.0509	0.0161	10.04	0.002	1.05	1.02	1.09
TIME_GAP	-0.0244	0.0062	15.61	0.000	0.98	0.96	0.99
AGE (20 & under)	0.6606	0.0596	122.97	0.000	1.94	1.72	2.18
AGE (21-25)	0.2867	0.0517	30.79	0.000	1.33	1.20	1.47
AGE (26-30)	0.1065	0.0538	3.93	0.048	1.11	1.00	1.24
AGE (31-35)	0.0317	0.0542	0.34	0.559	1.03	0.93	1.15
AGE (36-40)	0.0373	0.0542	0.47	0.491	1.04	0.93	1.15
AGE (41-45)	0.0186	0.0552	0.11	0.736	1.02	0.91	1.14
AGE (51-55)	0.0563	0.0606	0.86	0.353	1.06	0.94	1.19
AGE (56-60)	0.0700	0.0660	1.12	0.289	1.07	0.94	1.22
AGE (61-65)	0.1427	0.0737	3.74	0.053	1.15	1.00	1.33
AGE (66-70)	0.3013	0.0804	14.03	0.000	1.35	1.15	1.58
AGE (71-75)	0.6760	0.0864	61.26	0.000	1.97	1.66	2.33
AGE (76 & over)	1.1077	0.0788	197.38	0.000	3.03	2.59	3.53
GENDER (Male)	0.1216	0.0245	24.73	0.000	1.13	1.08	1.18
CR_TYPE (TYPE_2)	-0.0287	0.0268	1.15	0.284	0.97	0.92	1.02
CR_TYPE (TYPE_3)	1.4789	0.0349	1792.94	0.000	4.39	4.10	4.70
CR_TYPE (TYPE_4)	-1.6165	0.0872	344.01	0.000	0.20	0.17	0.24
Constant	-0.1760	0.0742	5.62	0.018	0.84		

Notes: Reference categories: SCHOOL - "1"; SUSPENSE - "1"; AGE - "46-50"; GENDER- "FEMALE"; CR\_TYPE - "TYPE\_1"

-2 Log likelihood – 40,372.938; Cox & Snell R Square - 0.125; Nagelkerke R Square - 0.167

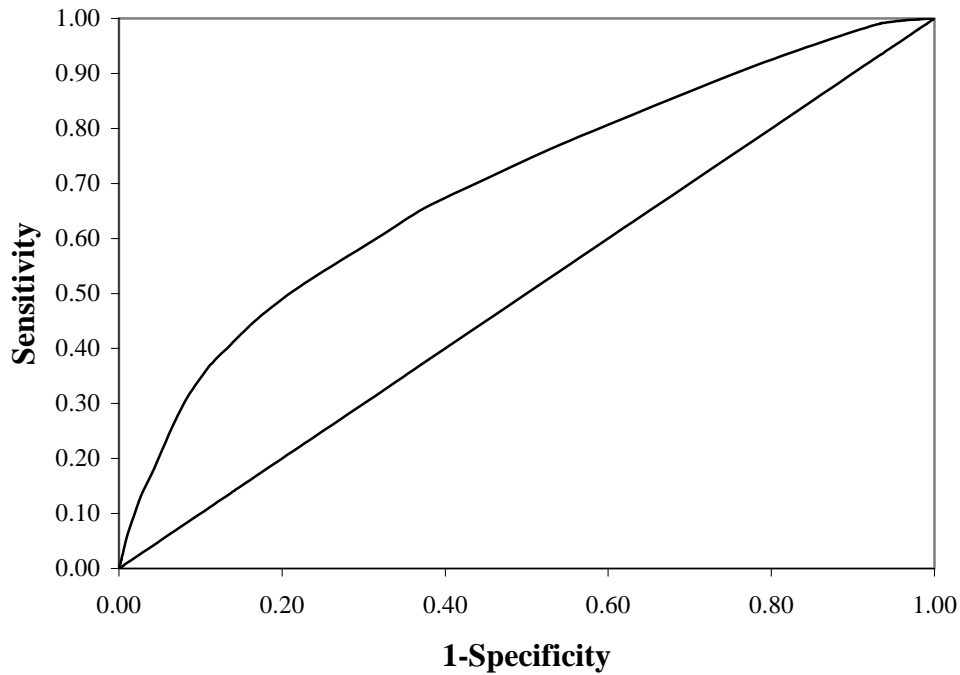
**Table 7: Result of logistic regression**

Removal Variable	df	-2 Log likelihood	Cox & Snell R <sup>2</sup>	Nagelkerke R <sup>2</sup>	AUC Value
None	n/a	40,372.94	0.125	0.167	0.6956
ATFAULT	1	40,467.26	0.123	0.164	0.6925
NOTFAULT	1	40,382.83	0.125	0.167	0.6953
SCHOOL	1	40,394.11	0.125	0.166	0.6948
SUSPENSE	1	40,385.03	0.125	0.167	0.6948
VIOLATE	1	40,382.71	0.125	0.167	0.6953
SPEEDING	1	40,383.02	0.125	0.167	0.6950
GAP	1	40,388.55	0.125	0.167	0.6960
AGE	5	40,862.32	0.113	0.150	0.6769
GENDER	1	40,397.67	0.124	0.166	0.6953
CR_TYPE	3	46,890.34	0.029	0.039	0.5947

Note: AUC values were estimated by using driver data which were not included in the calibration dataset.

**Table 8: Logistic regression model summaries**

The cut-point for the at-fault drivers for the model was established by compromising between sensitivity and specificity. However, the selected cut-point should produce reasonable low rates for false positive and false negative by the model because the Kentucky crash data can be randomly classified with almost a 50 percent chance of being correct (out of 69,155 drivers, 35,398 --51.2 percent-- were at-fault, 33,577 --48.6 percent-- were not-at-fault and the rest were not assigned a responsibility due to missing data). Figure 1 shows the Receiver Operating Characteristic (ROC) curve obtained for the data of 34,471 drivers who were not included in the calibration dataset. The observed area under the ROC curve is 0.6956, which would be generally accepted as a fair value in a 0.5 to 1 scale (32). This means that the model helps in classifying drivers' fault in a future crash as compared to a random guess. Thus, both overall efficiency and sensitivity were taken into consideration and two cut-points were suggested (Table 9). With the proposed model (Table 7), overall efficiency can be improved up to 64.15 percent with sensitivity of 65.54 percent and specificity of 62.51 percent when the cut-point is 0.48. However, it should be noted that the incorrect classification of not-at-fault drivers as at-fault drivers will not be as harmful as the opposite because all the drivers in question had multiple crashes. If this is the case, without decreasing specificity below the 50 percent level, the model can be used to improve sensitivity up to 74.56 percent, if a cut-point of 0.447 is used.



**Figure 2:** Receiver Operating Characteristic Curve for the best fit model

<b>Cut-point</b>	<b>Risk Predict</b>	<b>False positive Rate</b>	<b>False Negative rate</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>Overall Efficiency</b>
0.100	32,225	45.59	0.01	99.99	0.21	54.41
0.200	31,284	43.05	0.39	99.28	5.75	56.56
0.300	31,130	42.68	0.49	99.09	6.58	56.83
0.400	26,152	33.13	6.38	88.25	27.47	60.49
0.447	40,948	22.85	13.80	74.56	50.03	63.34
0.472	17,519	18.00	18.01	66.85	60.60	63.99
0.480	17,009	17.13	18.72	65.54	62.51	64.15
0.500	15,468	14.87	21.24	60.91	67.45	63.90
0.600	10,284	7.54	29.98	44.81	83.50	62.48
0.700	7,998	4.98	34.50	36.49	89.11	60.52
0.800	3,543	1.79	45.12	16.93	96.08	53.09
0.900	327	0.14	53.44	1.62	99.70	46.42

Note: Population 32,258; Missing 2,056; the total number of crash-involved drivers observed 17,522

**Table 9:** ROC trade offs for the best fit model for the experienced drivers

To determine the effect of each predictor variable on the drivers who had multiple crashes being at-fault in a future crash, odds ratios were used. If all other variables are kept constant in Eq. 1B, the odds ratio for a variable  $X_i$  can be estimated as  $\text{Exp}(B_i)$  where  $B_i$  is the coefficient of the variable  $X_i$ . Therefore, using Regression Coefficient in Table 7, odds ratios for the various variables are estimated and a summary is included in Table 7. For example, the odds ratio for ATFAULT is  $\text{Exp}(0.1536)=1.1660$ . In other words, a driver who had one previous at-fault crash is about 17 percent more likely to be at-fault in the next crash than a driver who had no previous at-fault crash involvements.



## 5.0 DISCUSSION AND CONCLUSIONS

According to the Kentucky crash data examined in this study, only a few drivers would be considered risky drivers compared to the total driving population. The data shows that only 4.93 percent of Kentucky's licensed drivers in 2002 were involved in a crash. However, over 43 percent of these drivers had experienced at least one or more previous crash incidents and 51.2 percent of drivers involved in these multiple crashes were at-fault. These at-fault drivers' actions caused not-at-fault drivers to be involved in crashes. Thus, identification of these at-fault drivers in advance assists licensing agencies in targeting them to take possible measures for improving overall traffic safety. Therefore, this study is aimed at determining crash prediction models with predetermined crash predictors that can be used to estimate the likelihood of a driver being responsible for a future crash occurrence.

Although the methodology adopted here is a modification of Gerber and Chen et al. studies, the major difference is that this study is focused *only* on determining the culpability in a crash that has occurred rather than determining the likelihood of a driver being involved in a crash (6,14). In addition, this study makes the maximum use of the latest available data by not dividing the crash period into two subdivisions and accounts for the effects of TIME\_GAP along with seven conviction categories.

It is widely accepted that crash prediction is difficult as is predicting the culpable driver (6,10,33). However, the model developed here could be used to identify as many as 74.56 percent of the at-fault drivers in 2002, if 0.447 of the cut-point is selected for the model. Moreover, this sensitivity was observed by using the holdout dataset which were not used in developing the model. This suggests that the methodology used here can be used to identify future high-risk drivers with higher confidence. However, for more accuracy, this should be tested with 2003 Kentucky data to establish the ability of the model to predict future at-fault drivers.

The ability to correctly identify 74.56 percent of at-fault drivers was achieved at the expense of the ability to correctly identify not-at-fault drivers. The model classifies these not-at-fault drivers

as at-fault drivers for the same reasons by which it correctly classifies at-fault drivers. It should be noted that most of these not-at-fault drivers were involved in crashes with the drivers who had no previous crash involvement and thus were not used in the analysis. Thus, the percentage of misidentified not-at-fault drivers as at-fault drivers by the model could be attributed to the potential misidentification of not-at-fault drivers although they were partially at-fault. As mentioned earlier, the prediction of not-at-fault drivers as at-fault drivers may not be as harmful as the opposite. The reason for this is that the prediction of at-fault drivers as safer drivers may prevent agencies and risky drivers from taking necessary steps to improve traffic safety. Identifying at-fault drivers in advance would allow agencies to take action to prevent such crashes or at least make such drivers aware of their potential risks. Moreover, results of such an effort will automatically eliminate some possibilities of innocent drivers being involved in crashes in which they were not at fault. Thus, this study recommends 0.447 for the cut-point of the model.

With respect to previous crash involvements, the positive coefficient for the ATFAULT of the model indicates that drivers who had previous at-fault crash(es) are more likely to be at-fault again in the event of a future crash. Similarly, as expected, the negative coefficient of NOTFAULT reveals that drivers who are more exposed to traffic hazards and/or are less defensive than the drivers in general, are more likely to be involved in crashes without being at fault. More research is needed to disaggregate NOTFAULT into two subgroups based on either drivers who were merely involved in crashes because of their higher exposure or drivers who are less defensive. A unit increase of ATFAULT increases drivers risk by 17 percent, whereas NOTFAULT reduces the risk by 5 percent if all other variables were kept constant. Moreover, drivers with an equal number of both at-fault and not-at-fault crashes in the past, still pose a risk of being at-fault in the future crash. However, one disadvantage of the model is that it tends to predict drivers as not-at-fault when they had more previous not-at-fault crashes than at-fault crashes. Thus, some form of data transformation before the model development would be an interesting subject for any future research.

NORISK, LAPSES and ERRORS are not significant at 0.05 level for the model determined here. Rather than the possible higher exposure, it is difficult to assume that any relation exists between

these NORISK drivers' crash involvements and the accumulation of NORISK convictions. However, this is not the case for LAPSES and ERRORS. First, it should be noted that there is a practical limitation to catch most drivers' LAPSES and ERRORS convictions compared to those of VIOLATE and SPEEDING. Second, both LAPSES and ERRORS are presumably not as results of drivers' intentional actions, and thus, drivers may gradually correct these problems themselves with time. However, as expected, VIOLATE and SPEEDING are significant predictors for the drivers being at fault in future crashes. The more VIOLATE or SPEEDING citations accumulated by a driver the riskier he/she will be. The model coefficients reveal that VIOLATE is riskier than SPEEDING. Each additional VIOLATE conviction increases a driver's chance to be at fault in the next crash by 7 percent, while SPEEDING makes it only 5 percent. However, it should be noted that this may be because of the generalization of conviction into these seven groups. No differentiation was made for the risk or severity of the citation within each group. For example, both 11-15 mph and 16-25 mph over-the-speed-limit convictions were received equal weight within SPEEDING , and both driving under the influence and reckless driving convictions received the same treatment within VIOLATE. The negative sign of the coefficient for SCHOOL and SUSPENSE indicates that drivers, who were treated by the current Kentucky Driver Point System for their past inappropriate driving behavior, had improved their driving abilities and they were less likely to be involved in crashes again as at-fault drivers.

The time gap between the most recent crashes is a good predictor for the drivers' crash responsibility in the model examined here. The results indicate that after an at-fault involvement the chance for a driver to be responsible for another at-fault crash is reduced with time. The advantage of the inclusion of this predictor is that the model allows the estimation of the future risk of the driver being responsible for a crash as a function of time after an at-fault crash involvement. Moreover, this variable allows identification of risky drivers who are more prone to be involved in crashes or who are highly exposed to more frequent traffic hazards.

The results of the study with respect to driver age and gender are consistent with those of prior research (6,21). The riskiest age group is the drivers over 76. However, they accounted for only 3.7 percent of the drivers in the analysis. The second riskiest group is the under 20 age group and they accounted fore 9.0 percent of the drivers in the analysis. Being young (age under 20) and

being older (age over 76) make them approximately 2 times and 3 times riskier than the age 46-50 group respectively. On top of this, being male makes drivers riskier than their female counterparts by 13 percent.

Although Table 8 shows that CR\_TYPE is the most important variable in the model, no significant difference was observed between intersection related crashes and non-intersection related crashes. Thus, the importance of the CR\_TYPE within the model should not overshadow the importance of other variables. However, drivers who are involved in TYPE 4 crashes are 4.39 times more likely to be designated as at-fault drivers by the police officers as compared to the drivers who were involved in crashes at intersections. On the other hand, drivers who are involved in TYPE 3 crashes are five times more likely to be designated as not-at-fault drivers as compared to the same reference category. It should be noted that all TYPE 3 and TYPE 4 crashes may be considered as single vehicle involved crashes. Thus, the assignment of the responsibility for these crashes depends mostly on the predetermined factors. For example, most of the drivers in the vehicles collided with animals were assigned as not-at-fault drivers in the Kentucky crash database.

As repeatedly mentioned, it should be noted that the model was developed only to predict drivers' responsibility if they are actually involved in a future crash. What is probably of most interest to licensing agencies is the ability to predict the likelihood of a driver being involved in a crash. Thus, more work is needed to improve the efficiency of driver control programs. However, once the likelihood of a driver being involved in a crash is estimated, this model will allow for estimating the drivers' chance of being at-fault in a future crash. Thus, it is of general interest to enhance the ability of the model presented here to estimate the drivers' probability of being at-fault in any future crash as well and explore the efficiency of the model using data from different jurisdictions.

In summary, it can be concluded that the model discussed here estimates with acceptable accuracy the probability of a driver's being at-fault in the event of a future crash occurrence. This may be a useful tool to determine at-fault drivers and thus, allow authorities to take actions aimed at improving overall traffic safety. The use of these models for renewing driver licenses

would be also another use of the model. The model can be also used for driver control programs aimed at preventing road crashes. These programs may vary from a simple form of issuing warning letters to a license suspension (i.e. the driver's license is temporarily withdrawn) or revocation (i.e. the driver's license is terminated). In addition to these two extremes, methods such as educational brochures, group educational meetings, diagnostic reexaminations, individual counseling, and administrative hearings can be considered.

## REFERENCES

1. Kentucky Traffic Collision Facts, Kentucky Transportation Center, College of Engineering, University of Kentucky, Lexington, Kentucky, 2002.
2. FHWA, Put the brakes on fatalities day, [http://safety.fhwa.dot.gov/fourthlevel/brakes\\_facts.htm](http://safety.fhwa.dot.gov/fourthlevel/brakes_facts.htm), 2001, accessed on Oct 13, 2003
3. Greenwood, M., Yule, U., An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, Vol. 83, No 2, 1920, pp. 255-279.
4. Blasco, R.D., Prieto, J.M., Cornejo, J.M., Accident probability after accident occurrence. *Safety Science*, Vol. 41, No 6, 2003, pp. 481-501.
5. Peck, R.C., Kuan, J., A statistical model of individual accident risk prediction using driver record, territory and other biographical factors, *Accident Analysis & Prevention*, Vol. 15, 1983, pp. 371-393.
6. Gebers, M.A., Strategies for estimating driver accident risk in relation to California's negligent-operator point system. California Department of Motor Vehicles Research and Development Branch, Technical Monograph 183, 1999.
7. Norris, F.H., Matthews, B.A., Riad, J.K., Characterological, situational, and behavioral risk factors for motor vehicle accidents: a prospective examination. *Accident Analysis and Prevention*. Vol. 32, 2000, pp 505-515.
8. Stamatiadis, N., Deacon, J.A., Trends in highway safety: Effects of an aging population on Accident Propensity. *Accident Analysis & Prevention*, Vol. 27, No 4, 1995, pp. 443-459.
9. Stamatiadis, N., Decon, J.A., Quasi-Induced exposure: Methodology and insight. *Accident Analysis and Prevention*, Vol. 29, No. 1, 1997, pp. 37-52.
10. Peck, R.C., McBride, R. S., Coppin R.S., 1971. The distribution and prediction of driver accident frequencies, *Accident Analysis & Prevention*, Vol. 2, pp. 243-249.
11. Mesken, J., Lajunen, T., Summala, H., Interpersonal violations, speeding violations and their relation to accident involvement in Finland, *Ergonomics*, Vol. 45, No 7, 2002, pp. 469 – 483.
12. Lui, K.J., Marchbanks, P.A., A study of the time between previous traffic infractions and fatal automobile crashes, 1984-1986. *Journal of Safety Research*, Vol. 21, 1990, pp. 45-51.

13. Hauer, E., Persaud, B. N., Smiley, A., Duncan, D., Estimating the accident potential of an Ontario driver. *Accident Analysis and prevention*. Vol. 23, No 2/3, 1991, pp. 133-152.
14. Chen, W., Cooper, P., Pinili. M., Driver Accident Risk in Relation to the Penalty Point System in British Columbia. *Accident Analysis and prevention*. Vol. 26, No 1, 1995, pp. 9-18.
15. Stamatiadis, N., Agent, K.R., Pigman, J., Ridgeway, M., Evaluation of retesting in Kentucky's driver license process. Research Report KTC-99-23, Kentucky Transport Cabinet, 1999.
16. Gebers, M.A., Peck, R.C., Using traffic conviction correlates to identify high accident-risk drivers. *Accident Analysis and Prevention* Vol. 35, 2003, pp. 903-912.
17. Parker, D., Reason, J. T., Manstead, A.S.R., Stradling, S.G. Driving errors, driving violations and accident involvement. *Ergonomics*, Vol. 38, no5, 1995, pp. 1036-1048
18. McKnight, A. J. and McKnight, A. S., Young novice drivers: careless or clueless? *Accident Analysis and Prevention* Vol. 35, No.6, 2003, pp. 921–925.
19. Kirk, A. and Stamatiadis, N. Traffic maneuver problems and crashes of young drivers. *Transportation Research Record* 1779, TRB, National Research Council, Washington, D.C., 2001, pp.68-74.
20. Kim, K., Li, L., Richardson, J., Nitz, L., Drivers at fault: Influences of age, sex, and vehicle type, *Journal of Safety Research*, Vol. 29, No. 3, 1998, pp. 171–179.
21. Chandraratna, S.K. and Stamatiadis, N., Problem Driving Maneuvers of Elderly Drivers. *Transportation Research Record* 1843, TRB, National Research Council, Washington, D.C., 2003, pp.89-95.
22. Massie, D.L., Green, P.E., Campbell, K.L., Crash involvement rates by driver gender and the role of average annual mileage. *Accident Analysis and Prevention*, Vol. 29, 1997, pp. 675–685.
23. Forward, S., Linderholm, I., Järmark, S., 1998. Women and traffic accidents, causes, consequences and considerations. In: *Proceedings of the 24th International Congress of Applied Psychology*, 9–14 August 1998. San Francisco.
24. Staplin, L., Lococo, K., Byington, S. and Harkey. D., Highway design handbook for older drivers and pedestrians. Federal Highway Administration, FHWA-RD-01-103, 2001.

25. Reason, J, Manstead, A, Stradling, S., Baxter, J., Campbell, K., Errors and violations on the road: a real distinction? *Ergonomics*, Vol. 33, No 10/11, 1990, pp. 1315-1332.
26. Peck, R.C., The identification of multiple accident correlates in high-risk drivers with specific emphasis on the role of age, experience and prior traffic violation frequency. *Alcohol, Drug & Driving*, Vol. 9 No 3-4, 1993, pp. 145-166.
27. Elliott, M.R., Waller, P.F., Raghunathan, T.E., Shope, J.T., Predicting offenses and crashes from young drivers' offense and crash histories. *Crash Prevention and Injury Control*, Vol. 2, No 3, 2001, pp. 167-178.
28. Johnson, D.E., *Applied multivariate methods for data analysts*. Duxbury press, Pacific Grove, 1998.
29. Norusis, M.J., 1999. *SPSS Regression Models 10.0*. SPSS, Inc., Chicago.
30. Hosmer, D.W. and Lemeshow, S. *Applied Logistic Regression*. John Wiley and Sons, New York, 1989.
31. Metz, C.E., ROC methodology in radiologic imaging. *Invest radiol*, Vol. 21, 1986, pp. 720-733.
32. Zweig, M.H., Campbell, G., Receiver-Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, Vol. 39, No 4, 1993, pp 561-577.
33. Raedt, D.R., and Kristoffersen, I. P., Predicting at-fault car accidents of older drivers. *Accident Analysis and Prevention*, Vol. 33, 2001, pp. 809-819.