Are novice drivers with crash history worse than other drivers?

N. Stamatiadis¹ & S. Chandraratna²
¹Professor, Department of Civil Engineering
University of Kentucky, USA
²Transportation Engineer, Woolpert LLP
Cincinnati, Ohio USA

Abstract

A goal for any licensing agency is the ability to identify crash-prone drivers. Thus, the objective of this study is the development of a crash prediction model that can be used to estimate the likelihood of a young novice driver being involved in a future crash occurrence. Multiple logistic regression techniques were employed using available Kentucky data. This study considers as crash predictors the driver’s total number of previous crashes, citations accumulated, and demographic factors. The driver’s total number of previous crashes was further disaggregated into the driver’s total number of previous at-fault and not-at-fault crashes. Sensitivity analysis was used to select an optimal cut-point for the model. The overall efficiency of the model is 77.82 percent and the model can be used to correctly classify more than a one third of potential crash-prone drivers if the cut-point of 0.247 is selected for the model. The total number of previous both at-fault and not-at-fault crash involvements, and having accumulated speeding citations are strongly associated with a driver being at risk. In addition, a driver’s risk is increased by being very young and being male. Although the statistical nature of driver crash involvements makes it difficult to accurately predict, the model presented here enables agencies to correctly identify 49.4 percent of crash-involved drivers from the riskiest top 500 driver group. Moreover, the model can be used for driver control programs aimed at preventing road crashes ranging from issuing warning letters to a license suspension.

1 Introduction

Several research studies have demonstrated that young drivers have crash rates significantly higher than those of any other groups (1,2). Their lack of driving experience is the most important reason behind their high crash risk (3). Even though young drivers’ crash risk still remained relatively high, numerous studies revealed that this risk fell dramatically after a few months of their licensure, and then gradually declined thereafter (3,4,5). This may be attributed more to an initial learning curve than to a rapid maturation (4). Harrison revealed that the most critical period is the first six months of a driving career (5). Moreover, Twisk observed a crash risk decrease for young drivers of almost two-thirds within the initial two-year period after they were licensed (6). This shows that categorizing each driver just based on their young age, even after a reasonable safe learning period, poses an unfair treatment to them. However, this is not the case for young drivers who had already been involved in a crash. Elliott et al. estimated that their previous-year crash involvements increase the risks of being involved in subsequent-year crashes by nearly 50 percent (2). Thus, this research made attempts to determine a model which can be used to identify Kentucky young novice drivers who may pose a prolonged threat to public safety. To this end, multiple logistic regression techniques were employed for the subjects in order to establish relationships between these drivers’ potential crash involvements and crash predictors. The model presented here enables licensing agencies to develop possible remedies and alert risky young novice drivers of their potential to be involved in a crash.

2 Background

The goal for any licensing agency is the ability to identify high-risk drivers. 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 (7). During the1999 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 Kentucky crashes in 2002 resulted in the loss of $5.9 billion worth of quality of life associated with deaths and injuries as well as other economic costs (7). The national figures are more alarming. In 2001, the 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 (8). Moreover, each year more than 40,000 people are killed and more than 5 million are injured on U.S. highways.

There is always some degree of risk of being involved in a crash whenever one drives a motor vehicle but human error is the single greatest contributing factor in most road crashes and is more notable when drivers’ recurrent crashes are examined. After analyzing a Spanish bus drivers’ crash database during 1976-1983, Blasco et al. concluded that these drivers’ recurrent crashes primarily occurred because of human error rather than by chance (9). Literature shows that not
only previous crash involvements, but also accumulations of citations are good predictors of drivers’ future crash risk. Hauer et al. 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 (10). Moreover, many researchers repeatedly highlighted the importance of past crash and citation records in predicting drivers’ potential crash involvements (11,12,13).

In response, agencies throughout the world use driver control systems as a deterrent to negligent driving in order to prevent crash involvements. These programs vary from issuing warning letters to license suspension (i.e. the driver’s license is temporarily withdrawn) to license revocation (i.e. the driver’s license is terminated). Currently, the Division of Driver Licensing of the Kentucky Transportation Cabinet uses a Point System to identify drivers who are so-called habitually negligent drivers (14). Under the Point System, each driver starts with no points, but accumulates points as a result of various offenses. Different points are assigned to each conviction based on the relative importance of the action as deemed by the Kentucky State Police. If a driver accumulates 12 points or more (7 points or more if under age eighteen) within a two year period, a driver's privilege to operate a motor vehicle may be suspended.

These driver control programs not only incur substantial costs to agencies but also have the potential to affect an individual’s freedom to travel. Examining 362,053 Kentucky drivers who were involved in crashes during 2001-2002 period, it was observed that 94,351 of them (26.1%) had a crash even after they were subjected to either post-license control or driver improvement programs (e.g. traffic school). On the other hand, of the 362,053 drivers, 228,261 (63.0%) of them had maintained a citation-free record during the 1995-2000 period. Thus, it is important that the drivers targeted for these programs are identified correctly and accurately. To this end, a substantial body of research has specifically focused on improving the ability to identify high-risk drivers using their past crash and citation records as mentioned earlier (11,12,13,15).

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 the California Department of Motor Vehicles 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 (16). For example, Gerbers’ model, which was determined by logistic regression using age, gender, license class, 17 types of citations and total number of crash involvements, can be considered (12). Examining this model’s results, it was observed that a 72.4 percent of crash-free drivers were erroneously classified to be crash-involved and a similar percentage of crash-involved drivers were erroneously classified to be crash-free. Moreover, deploying canonical correlation in another more recent research, Gerbers and Peck noticed that 72.8 percent of the drivers predicted to be crash-involved remained crash-free by their best model (13). Thus, an adoption of such a model for a driver control program may penalize a significant number of innocent drivers. However, it should be noted that most of these studies targeted the complete driver population and were not limited to a certain risky group with similar characteristics.

3 Methodology

Multivariate statistical procedures allow the development of a model for a driver population representing different groups with different characteristics. However, it is reasonable to assume that targeting a high-risk driver group rather than the entire driver population helps to improve a risk model’s efficiency. Even though such a model cannot be used for an entire driver population, an ability to identify risky drivers more accurately should not be underestimated because correct identification is important for better treatments. Taking into account young novice drivers’ problematic behaviors, this study was only limited to Kentucky young novice drivers aiming at determining a crash model which produces less erroneous hits than the other existing models. The “Kentucky young novice drivers” refers here to the drivers who were under age 25 and had their license for exactly 2 years.

Two databases were used in this analysis for the 1997-2002 period. The Kentucky Driver License (KyDL) database was used to identify the drivers to be used in this study and complete records for these drivers were obtained for the 1997-2002 period. A total of 3,201,620 driver records were used in this analysis. However, the KyDL database does not provide a detailed crash history and the Kentucky Crash (KyC) database was used to identify the details for each crash of the Kentucky drivers for the 1997-2002 period. For crashes with more than one driver involved, separate records were created, since the unit of analysis in this study is each licensed driver.

The two databases were merged by matching the driver’s license number. However, a slightly different procedure was adopted to select data for the study here from the similar past studies. In the past studies, the study period was subdivided into two intervals in order to make a prediction of subsequent crash risk during a predetermined period based on data collected over a fixed predictor period. For example, Gerbers used the first four-year (i.e. 1984-1987) complete record of crash involvements and conviction records to identify the high-risk drivers who are most likely to be involved in crashes in the following four years, which was called subsequent period (i.e. 1988-1991). 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, this study was limited to young novice drivers who had exactly completed two-years of driving career.
during the predictor period. As noted before, this is the most influential period for crash risk \((4).\) Thus, a selection of fixed two periods would not allow for having a reasonable number of subjects for the study sample. As an alternative, a floating two-year predictor period was selected for each driver starting from the day on which the driver was licensed. Then, the subsequent two-year period was adjusted accordingly. This procedure produces a data set of 64,562 driver subjects with two-year predictor and subsequent periods during the 1997-2002.

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 \((17)\). 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. However, most past studies have considered drivers’ total number of crashes or total number of at-fault crashes as a predictor variable on predicting the individual driver’s likelihood of being involved in future crashes. This procedure created some methodological problems for the models because they are determined in predicting not only future at-fault involvement but also not-at-fault crash involvement. Thus, both the number of at-fault (AFault) and not-at-fault crashes (NOTFAULT) were considered as independent predictor variables here.

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 \textit{“no liability insurance in force.”} Therefore, such citations, along with all other non-moving violations, were ignored for this study. All other moving violations were considered as risky behavioral violations and were categorized into two groups -- non-speeding violations (NONSPEED) and speeding violations (SPEEDING). 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. In addition, these drivers were considered as treated ones by the current driver control program. So were those whose licenses were suspended and who had received warnings. It should be noted that it is desirable to develop a simple model form to identify risky drivers to ease model implementation. However, the literature suggests that this may be an important variable and it was included here but without considering the specific types for treatment to minimize the different combinations of program types (e.g. Suspension and Traffic School, Suspension and Warning Letters, and so forth). The inclusion of such combinations makes the model more complicated. Thus, to test the effects of the current programs, this information was gathered in to one group (TREATED) and was included in the analysis.

As pointed out by the large body of research, results of a recent study on Kentucky young drivers also showed that there is a general trend of decreasing crash involvement risk with increasing age \((3)\). Similarly, gender of the driver has also been used in the past as a good crash predictor \((12)\). Thus, both age (AGE) and gender (GENDER) of drivers were examined here as well. All variables were treated as continuous variables except TREATED and GENDER which are categorical variables.

Logistic regression has proven to be the most appropriate statistical technique for this type of modeling \((2,12,17)\). The reason is that logistic regression is a form of regression which is used when the dependent variable is binary; that is, it can have only two values (crash occurrence or not) \((18,19)\). The 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 require normally distributed variables, and in general has less stringent requirements than linear regression. In this analysis, the dependent variable is the future crash involvement of the driver. Thus, the probability of occurrence of a crash is modeled as follows:

\[
\text{Prob(being involved in a future crash)} = \frac{1}{1 + e^{-z}}
\]

\(\text{(Eq.1)}\)

Where \(Z\) is the linear combination

\[Z = B_0 + B_1X_1 + B_2X_2 + \ldots + B_NX_N\]

and the B’s are coefficients, estimated using the maximum-likelihood method, and the X’s are the independent predictor variables previously discussed.

An advantage of models derived by logistic regression, besides the ability to predict the probabilities of drivers being involved in a crash, 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. 2).

\[
\frac{\text{prob(being involved in a future crash)}}{\text{prob(being not involved in a future crash)}} = e^{B_0 + B_1X_1 + \ldots + B_pX_p}
\]  
(Eq. 2)

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. SPSS statistical software was used throughout this study and SPSS’s select random sample of cases procedure was employed to create two datasets -- a calibration dataset for model development and a holdout data set for the model validation. This is vital because the usefulness of a risk model depends on how correctly it can identify high-risk drivers from the data which were not used in developing the model. Because of the large sample size the holdout data procedure will not adversely affect developing the best model (18).

4 Results

4.1 Statistical Analyses

Out of 64,562 subject drivers, 32,205 drivers’ data were randomly selected for the calibration data set while the remaining 32,357 drivers’ data were kept in the holdout data set for the model validation. Then, logistic regression was adopted to determine the best fit model between the dependent variable and the independent variables using the calibration data set. The dependent variable was coded to 0 if there is no crash involvement and 1 if there is one or more crash involvement observed during the subsequent period. All the coefficients of the model were tested based on the Wald Statistic. It was observed that five independent variables -- ATFAULT, NOT FAULT, SPEEDING, AGE, and GENDER --were significant at 0.05 level for this model whereas the variables TREATED (p=0.114) and NONSPEED (p=0.831) were not. The summary of the model is given in Table 1.

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>RegCoef</th>
<th>Standard Error</th>
<th>Wald Statistic</th>
<th>p-value</th>
<th>Odds Ratio</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATFAULT</td>
<td>0.9147</td>
<td>0.0250</td>
<td>1335.53</td>
<td>0.0000</td>
<td>2.50</td>
<td>2.38 - 2.62</td>
</tr>
<tr>
<td>NOTFAULT</td>
<td>0.8095</td>
<td>0.0311</td>
<td>677.65</td>
<td>0.0000</td>
<td>2.25</td>
<td>2.11 - 2.39</td>
</tr>
<tr>
<td>TREATED</td>
<td>0.0748</td>
<td>0.0474</td>
<td>4.29</td>
<td>0.1143</td>
<td>1.08</td>
<td>0.98 - 1.18</td>
</tr>
<tr>
<td>NONSPEED</td>
<td>0.0050</td>
<td>0.0236</td>
<td>0.05</td>
<td>0.8305</td>
<td>1.01</td>
<td>0.96 - 1.05</td>
</tr>
<tr>
<td>SPEEDING</td>
<td>0.1938</td>
<td>0.0307</td>
<td>39.86</td>
<td>0.0000</td>
<td>1.12</td>
<td>1.11 - 1.29</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.2014</td>
<td>0.0037</td>
<td>19.86</td>
<td>0.0000</td>
<td>1.15</td>
<td>1.08 - 1.23</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.1303</td>
<td>0.0116</td>
<td>127.23</td>
<td>0.0000</td>
<td>0.88</td>
<td>0.86 - 0.90</td>
</tr>
<tr>
<td>Constant</td>
<td>0.0744</td>
<td>0.2020</td>
<td>0.14</td>
<td>0.7125</td>
<td>1.08</td>
<td></td>
</tr>
</tbody>
</table>

Reference categories: TREATED - “Not treated (0)”; GENDER- “FEMALE (0)”-2 Log likelihood – 27,103.827; Cox & Snell R Square - 0.076; Nagelkerke R Square - 0.126

One of the advantages of logistic regression is the ability to determine the effect of each predictor variable on the drivers’ chance of being involved in future crashes by using odds ratios. These ratios allow for the estimation of the relative change on a driver’s crash involvement risk if only the variable in question is increased by a unit. Therefore, keeping all variables but one constant in Eq. 2, the odds ratio for this variable (X_i) will be Exp(B_i) where B_i is the coefficient of the variable. The regression coefficients in Table 1 allow for estimating the odds ratios for each of the predictor variables, which are also shown in Table 1. For example, the odds ratio for SPEEDING is Exp(0.1938)=1.21 which indicates that if a driver receives one more speeding citation the chance of being involved in a crash increase by 21 percent. Table 1 reveals that a unit increase of all the significant variables, with the exception of AGE, will increase individuals’ risk to be involved in a future crash.

With respect to previous crash involvements, the positive coefficient for the ATFAULT of the model indicates that drivers who had previous at-fault crashes are more likely to be involved in additional crashes than the rest of the drivers. A
driver who had one previous at-fault crash is about 150 percent more likely to be involved in a recurrent crash than a driver who had no previous at-fault crash involvements. Similarly, as expected, the positive 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 recurrent crashes as well. A unit increase of NOTFAULT increases drivers risk by 125 percent, which is slightly less than TOFAULT risk effects. It can be seen that these risks are much higher than the Elliott et al. findings, although both studies had similar interests with only few exceptions (2). Elliott et al. didn’t consider drivers’ total number of not-at-fault crashes as a predictor variable and this study was limited to young drivers with exactly two years of experience after their license.

More surprisingly, TREATED is not significant at the 0.05 level for the model determined here. This raises a question about the effectiveness of existing driver control programs on young novice drivers. Even after the drivers are “treated” by the current Kentucky Driver Point System for their past inappropriate driving behavior, they still cannot be considered safe drivers. However, this may be due to the fact that traffic school and suspensions were combined into a single TREATED group. It is possible that each of these treatments may have a different effect on the drivers’ propensity to be involved in a crash as past research revealed that suspensions have positive effects on future crashes while Traffic Schools do not (20). Thus, the separation of the treatments to unique categories may improve the accuracy of the model and it will be tried as part of the continuing research in this area.

NONSPEED is not a significant crash predictor either, however, it should be noted that this may be because of the generalization of conviction into only two groups -- NONSPEED and SPEEDING. No differentiation was made for the risk or severity of the citation within the group. For example, driving under the influence and reckless driving convictions received the same treatment within NONSPEED. However, as expected, SPEEDING is a significant predictor for the drivers being at risk. The more speeding citations accumulated by a driver, the riskier he/she will be. An additional speeding conviction increases a driver’s chance to be involved in a crash by 21 percent. As for NONSPEED, no account was made for the high speed level differences as well and for example, both 11-15 mph and 16-25 mph over-the-speed-limit convictions received equal weight. Thus, it would be interesting to see the effects of further desegregation of these two conviction groups.

The results of the study with respect to driver age and gender are consistent with those of prior research (3,12). The results indicate that the very young have an increased chance of being involved in future crashes. The negative sign of the coefficient for AGE indicates that young drivers’ future risk decreases by an average of 12 percent per year from one year to the next. On top of this, being male makes drivers riskier than their female counterparts by 15 percent.

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 safe drivers and drivers with values greater than the cut-point as risky drivers. Thus, it is understood that the number of drivers predicted as risky and safe depends on the cut-point selected by the analysis. However, the usefulness of a model depends on how many more risky drivers it can accurately classify than would be expected by chance alone. The Receiver Operating Characteristic (ROC) curve is considered to be the most common global measure to quantify the diagnostic accuracy of a model (21). ROC curves provide a pure index of accuracy by demonstrating the limits of the model’s ability over the complete range of cut-points. The observed area under the ROC curve was 0.712 which means that a randomly selected individual from the risky driver group has a predicted probability larger than that for a randomly chosen individual from the safer group 71.2% of the time. This indicates that the model clearly helps in classifying drivers’ crash proneness as compared to a random guess (21).

4.2 Selection of Cut-point

The cut-point for the model can be established by compromising between sensitivity (i.e. the proportion of the crash-involved drivers that can be correctly predicted) and specificity (i.e. the proportion of crash-free drivers that can be correctly predicted) by minimizing the two error values for the false negatives and false positives to an acceptable level. Table 2 shows how model properties change with the cut-point. Unfortunately, as it can be seen, these two erroneous classifications are reciprocally related. Any decrease of the false negative predictions causes an increase in the false positive predictions. Thus, a 0.247 cut-point, which produced an equal number of false negatives and false positives with a sensitivity of 35.92%, was selected to identify risky drivers and make the misclassification rates the same.
### TABLE 2 Diagnostic test results with respect to cut-point

<table>
<thead>
<tr>
<th>Cut point</th>
<th>Number of high-risk drivers classified</th>
<th>True positive</th>
<th>False negative</th>
<th>False positive</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Over all Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.100</td>
<td>26,994</td>
<td>5,266</td>
<td>335</td>
<td>21,728</td>
<td>94.02</td>
<td>18.79</td>
<td>31.81</td>
</tr>
<tr>
<td>0.200</td>
<td>10,014</td>
<td>3,378</td>
<td>2,223</td>
<td>6,636</td>
<td>60.31</td>
<td>75.20</td>
<td>72.62</td>
</tr>
<tr>
<td>0.247</td>
<td>5,601</td>
<td>2,012</td>
<td>3,589</td>
<td>3,589</td>
<td>35.92</td>
<td>86.59</td>
<td>77.82</td>
</tr>
<tr>
<td>0.300</td>
<td>3,057</td>
<td>1,142</td>
<td>4,459</td>
<td>1,915</td>
<td>20.39</td>
<td>92.84</td>
<td>80.30</td>
</tr>
<tr>
<td>0.400</td>
<td>2,206</td>
<td>833</td>
<td>4,768</td>
<td>1,373</td>
<td>14.87</td>
<td>94.87</td>
<td>81.02</td>
</tr>
<tr>
<td>0.500</td>
<td>659</td>
<td>284</td>
<td>5,317</td>
<td>375</td>
<td>5.07</td>
<td>98.60</td>
<td>82.41</td>
</tr>
<tr>
<td>0.600</td>
<td>420</td>
<td>182</td>
<td>5,419</td>
<td>238</td>
<td>3.25</td>
<td>99.11</td>
<td>82.52</td>
</tr>
<tr>
<td>0.700</td>
<td>145</td>
<td>70</td>
<td>5,531</td>
<td>75</td>
<td>1.25</td>
<td>99.72</td>
<td>82.67</td>
</tr>
<tr>
<td>0.800</td>
<td>56</td>
<td>29</td>
<td>5,572</td>
<td>27</td>
<td>0.52</td>
<td>99.90</td>
<td>82.70</td>
</tr>
<tr>
<td>0.900</td>
<td>5</td>
<td>2</td>
<td>5,599</td>
<td>3</td>
<td>0.04</td>
<td>99.99</td>
<td>82.69</td>
</tr>
</tbody>
</table>

Note: Population 32,357; the total number of crash-involved drivers observed in the holdout dataset -5,601

### 4.3 High risk drivers

In addition to selecting a cut-point, the driver population in question was pooled into several groups of drivers with respect to their risk level. In other words, the top 100, 500, 1000, 5,000 and 10,000 drivers who had the highest probability of being involved in crashes were selected and then the drivers who were actually involved in crashes during the subsequent period were counted. These findings are listed in Table 3 and indicate that the smaller the risky driver pool is, the better the model is at classifying crash-prone drivers. For example, if the model’s high-risk driver group is limited to the 100 drivers of the driver population in question with the highest chance to be involved in a crash, 50 of them were correctly assigned to the group (i.e. these 50 drivers out of 100 drivers were actually involved in crashes during the subsequent two-year period).

### TABLE 3 Identification of risky drivers with two years of driving experience

<table>
<thead>
<tr>
<th>Designated risk group</th>
<th>Drivers with crashes</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>50</td>
<td>50.0%</td>
</tr>
<tr>
<td>500</td>
<td>247</td>
<td>49.4%</td>
</tr>
<tr>
<td>1,000</td>
<td>444</td>
<td>44.4%</td>
</tr>
<tr>
<td>5,000</td>
<td>1,906</td>
<td>38.1%</td>
</tr>
<tr>
<td>10,000</td>
<td>3,696</td>
<td>37.0%</td>
</tr>
</tbody>
</table>

### 5 Discussion and conclusions

The importance of improving the knowledge regarding factors that increase risk for vehicle crashes is undisputed if the crash facts are closely observed. Reviewing past studies, it can be hypothesized that some drivers have a higher likelihood of being involved in crashes than can be expected by chance, so they can be considered high-risk drivers. However, the degrees of precision of the available risk models to identify them are still not high, although a substantial effort has been made in improving them. Elliott et al. suggests that this may be because these studies were based on the complete driver population with different groups with different characteristics. Therefore, the subjects for this study were limited to the Kentucky young novice drivers group with two-year driving experience. As expected, a considerable improvement was observed in the model’s performance. The model could be used to identify as many as 35.92 percent of the crash-prone drivers from the subject driver group during the subsequent two-year period, if 0.247 of the cut-point is selected for the model.

Although the methodology presented here can be considered an extension of previous studies, it should be noted that all the conclusions here were made on the results obtained by using the holdout data set which was 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 the future Kentucky data to establish the ability of the model to predict future crash-involved drivers.
The model presented here shows that the total number of previous crash involvements (both at-fault and not-at-fault), accumulated speeding citations, age, and gender are strong predictors of a driver’s crash risk. However, non-speeding citations and whether a driver received a corrective action (traffic school or license suspension) were not significant for this model. As previously noted, this may be mostly due to the lack of separating the corrective actions. Additional work is likely needed to further refine the variables used to determine their specific potential contribution to predict crash risk. However, this has the potential to increase the number of independent variables in the model and may make the model more complicated. Considering the model efficiency, it is of general interest to enhance the ability of the model presented here to estimate the young novice drivers’ probability of being involved in crashes. Other than further desegregation of conviction groups, some form of data transformation before the model development would be an interesting subject for any future research. In this study, all the data were used as they were counted from the databases. However, it is well known that sometimes transformation of data values such as the total number of speeding convictions by means of some mathematical function, such as square root of the original value, may yield datasets that may have better statistical properties than the original data.

It should be noted that the subjects considered here can not be considered mature drivers. These models may be a useful tool in identifying risky drivers from the driver group in question and thus allow authorities to take actions aimed at improving overall traffic safety. The models can also be 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 or revocation. Masten and Peck has pointed out that “License control actions by far are the most effective countermeasures; license suspension should be triggered as soon as is legally feasible (20).” However, they did not rule out the use of warning letters and pointed out how an appropriate level or duration of the program can be selected based on an individuals’ risk level. For example, since 1987 the California program has used a four level approach based on points accumulated with: (a) a soft advisory letter (Level 1); (b) a hard-threat warning letter (Level 2); (c) short license suspension plus license probation (Level 3); and (d) probation violator license revocation (Level 4). In between levels 1 and 4, other methods such as educational brochures, group educational meetings, diagnostic reexaminations, individual counseling, and administrative hearings have been recommended as alternatives aiming at improving traffic safety (20). The current study shows that it is not possible to perfectly identify risky drivers from the driver population. For example, 5,601 drivers were predicted as risky but only 35.92% of them were actually involved in crashes during the subsequent period. Thus, warning letters are the best solution for these risky drivers because of their low-cost and large-volume viability and to minimize penalizing innocent drivers. On the other hand, it was observed that 247 (49.4%) drivers out of 500 drivers of the population with the highest chance to be involved in a crash were actually involved in crashes during the subsequent period. Thus, the best solution for them is license suspension as Masten and Peck revealed (20).

In summary, it can be concluded that the models discussed here estimate with acceptable accuracy the probability of a young novice driver being involved in a future crash. However, additional work may be needed to further enhance our ability to identify risky drivers and determine more accurately the most appropriate predictors for this task.

REFERENCES