STATE OF CALIFORNIA
DMV
DEPARTMENT OF MOTOR VEHICLES

STRATEGIES FOR ESTIMATING DRIVER ACCIDENT RISK IN RELATION TO CALIFORNIA'S NEGLIGENT-OPERATOR POINT SYSTEM

By
Michael A. Gebers

July 1999

California Department of Motor Vehicles
Research and Development Branch
Technical Monograph 183

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
July 1999

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Strategies for Estimating Driver Accident Risk in Relation to California's Negligent-Operator Point System

5. FUNDING NUMBERS

6. AUTHOR(S)
Michael A. Gebers

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
California Department of Motor Vehicles
Research and Development Section
P.O. Box 932382
Sacramento, CA 94232-3820

8. PERFORMING ORGANIZATION REPORT NUMBER
CAL-DMV-RSS-99-183

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

A sample of approximately 140,000 records of licensed California drivers containing information on age, gender, and driving record variables was examined. The goal of this paper was to assess the accuracy of predicting future accident risk using various combinations of demographic and prior driving record variables as predictors in 17 regression models.

All of the models were consistent in demonstrating that increased probability of subsequent accident involvement is associated with increased prior citation and prior accident frequencies, being young, and being male. Results from the regression models indicated the following:

- Models that use prior total accidents as a predictor variable perform better than models that do not use prior total accidents as predictors.
- Models that use prior culpable accidents as a predictor do not perform as well as models that use prior total accidents as a predictor.
- A comparison of models in which 17 individual violation types are used as predictors to those in which only total citations is used as a predictor shows only a small advantage of using individual violation types.
- Models that use as predictors the demographic variables of age, gender, and license class along with various combinations of citations and accidents perform better than California’s current neg-op system, which uses a weighted combination of countable citations and responsible accidents.

It was concluded that if the goal of driver record adjudication systems is to identify and apply sanctions to high-risk drivers in order to intervene before this risk is realized, then the results presented in this report support the current point-count strategy which attempts to optimize the identification of drivers having a high probability of subsequent accident involvement.

14. SUBJECT TERMS
Motor vehicle accidents, traffic safety, accident proneness, accident rates, accident risks, accident repeater drivers, convictions, high-risk drivers

15. NUMBER OF PAGES
51

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
None

PREFACE

This report is issued as an internal monograph of the California Department of Motor Vehicles' Research and Development Branch rather than an official report of the State of California. The opinions, findings, and conclusions expressed in the report are those of the author and not necessarily those of the State of California.

ACKNOWLEDGMENTS

This study was conducted under the general direction of Raymond C. Peck, Research Chief, and the supervision of Robert A. Hagge, Research Manager.

The author would like to extend thanks to Dr. Mary Janke, Acting Research Chief, for reviewing and editing of the report. The author would also like to thank Douglas Luong, Management Services Technician, for typing drafts of the report, tables, and figures, and Debbie McKenzie, Associate Governmental Program Analyst, for assisting in the preparation and proofreading of the report and coordinating its production.

EXECUTIVE SUMMARY

Background

- Driver control systems assign penalty points to various traffic law infractions and establish a level of point accumulation at which a licensing action is taken. These driver control systems are used by almost all motor vehicle departments.

- The primary objective of point systems is to identify and initiate driver improvement or license control actions against drivers most likely to become accident involved. The existence of point systems presumably serves as a general deterrent to negligent driving and to the accumulation of numerous traffic citations.

- In California's negligent-operator (neg-op) point system, each conviction of a violation of the traffic law carries a specific number of neg-op points. Following conviction of the offense, these points are added to the individual's driving record. When the point count reaches specified levels, the driver is exposed to a neg-op "treatment." This treatment usually takes the form of a warning letter for the lowest specified neg-op point level and climaxes in the suspension or revocation of the driver license at the highest level of neg-op action.

- Scientific literature indicates that it is difficult to accurately identify individual accident-prone drivers on the basis of their prior traffic accident and conviction records. But a large body of research has established that statistically significant relationships exist between counts of traffic accident involvements and counts of prior traffic accidents and citations for groups of drivers. The present paper further explores accident-risk modeling by building on the techniques presented by Chen et al. (1995) and Hauer et al. (1991).

Study Objective

- The analyses presented in this paper are designed to estimate the following through the application of a variety of accident prediction models:

1. A driver's accident risk, defined as the expected probability of accident involvement during the subsequent accident criterion period.

2. The accuracy of prediction, as measured by the rate of true positives (i.e., drivers correctly predicted to be accident-involved) and true negatives (i.e., drivers correctly predicted to be accident-free).

- It is intended that the results presented in the paper will be used to estimate the accident risk levels of identifiable subgroups in the driver population and to assist in the ongoing refinement of California's point system and other safety programs for selecting negligent and accident-prone drivers for treatment and driver control actions.

Research Methods

- Data for the analyses were obtained from the driving records of approximately 140,000 licensed drivers from a 1% random sample of the California driving population, extracted in 1992 from the California Driver Record Study Database.

- For each subject, information was collected on demographic factors like age and gender and driver record information such as total accidents, total citations, responsible accidents, neg-op points, and individual violation types (e.g., speeding, right-of-way, and running red light).

- Multiple logistic regression was used to develop and assess a number of prediction models. Specifically, estimated prediction models were developed to compare:

1. Models that use age, gender, and license class as predictor variables versus models that do not.

2. Models that use at-fault or "responsible" accidents among the set of predictor variables versus models that use total accidents among the set of predictor variables.

3. Models that include separate parameters for 17 individual violation types as predictors versus models that have one common parameter for all citation types combined.

4. Models that use the number of 0-, 1-, and 2-point citations as predictors versus models that do not.

5. Models that use the number of failure-to-appear (in court; FTA) violations as a predictor versus models that do not.

6. Models that use the number of traffic violator school (TVS) dismissals as a predictor versus models that do not.

7. Models that use the number of neg-op points as a predictor versus models that do not.

- The following table summarizes the models and variables that were evaluated.

Predictor Variables Evaluated in Each Model

Model	Demographics (age, gender, license class)	Citations		Accidents		FTA	0, 1, & 2 points	Neg-op points	TVS dismissals
		17 types	Total	Responsible	Total				
A1	X	X							
A2	X	X			X				
A3	X	X		X					
B1		X							
B2		X			X				
B3		X		X					
C1	X		X						
C2	X		X		X				
C3	X		X	X					
D1			X						
D2			X		X				
D3			X	X					
E1	X		X		X	X			
E2	X		X	X		X			
F	X				X		X		
G								X	
H	X		X		X				X

Note. An X indicates the inclusion of the variables in the model.

The models were evaluated using a number of different techniques to determine the following:

Which model is best in identifying a driver's accident risk, as defined by the predicted probability of accident involvement, and manifests the highest level of predictive accuracy in discriminating accident-involved from accident-free drivers?

Results

Model comparisons. Parameter estimates were computed for the 17 logistic regression models defined in the above table. Results from the regression models indicated the following:

- Models that use prior total accidents as a predictor variable perform better than models that do not use prior total accidents as a predictor.

- Models that use prior culpable accidents as a predictor do not perform as well as models that use prior total accidents as a predictor.

- A comparison of models in which the 17 individual violation types are used as predictors to those in which only total citations is used as a predictor shows only a small advantage of using the individual citation types.

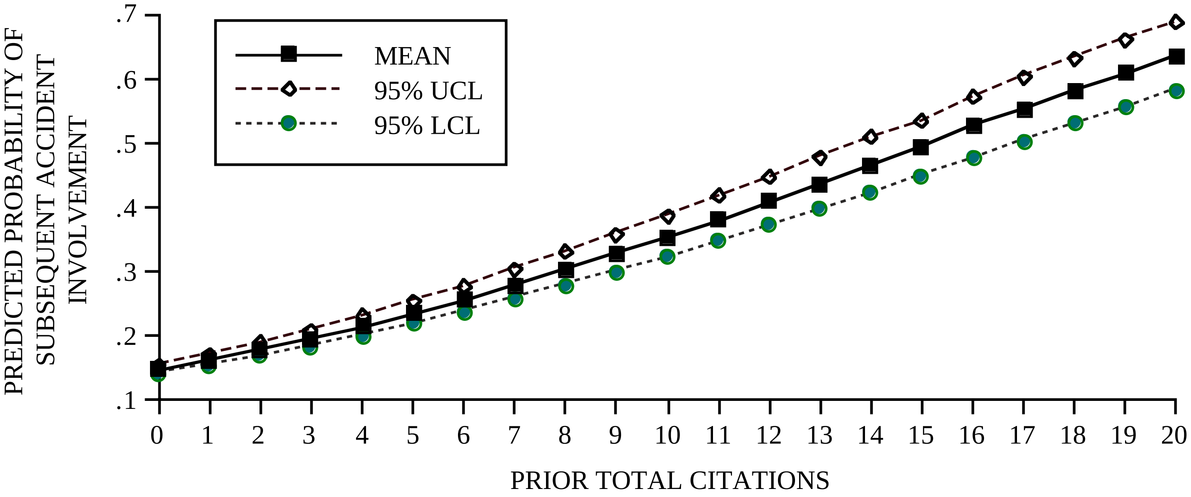
- Models that use as predictors the demographic variables of age, gender, and license class along with various combinations of citations and accidents perform better than

California's current neg-op system, which uses a weighted combination of countable citations and responsible accidents.

An examination of the logistic regression parameters indicated that models F and H produced the best fit. Model F consists of age, gender, license class, the number of total accidents, and the number of 0-, 1-, and 2-point citations as predictors. Model H consists of age, gender, license class, prior total citations, prior total accidents, and the number of TVS dismissals as predictors. Results from these two models indicate that increased probability of accident involvement is associated with:

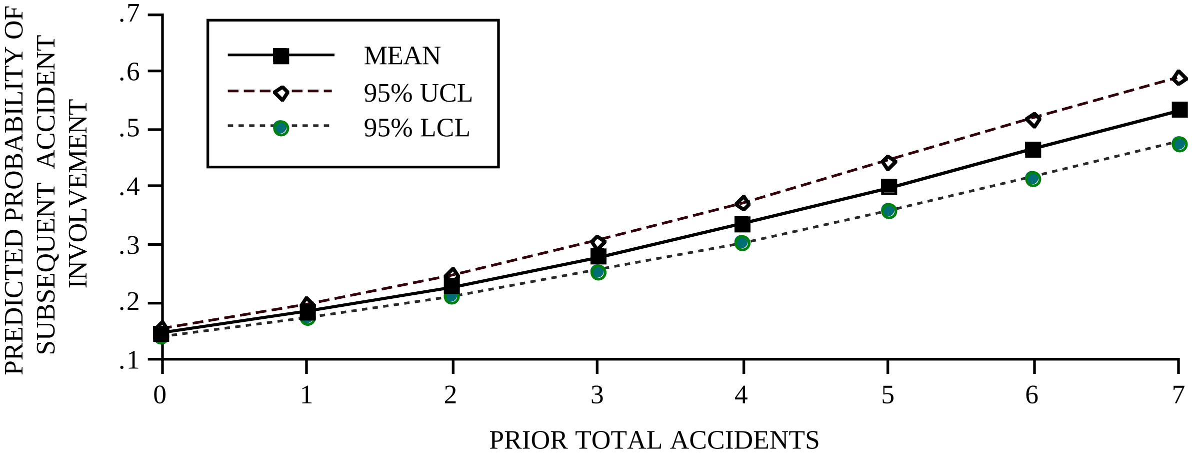
- Being young
- Being male
- Holding a commercial driver's license
- Increased prior citation frequency
- Increased prior accident frequency

The figure below displays the predicted probability (mean) and 95% confidence intervals (upper confidence limits - UCL and lower confidence limits - LCL) for subsequent 4-year total accident involvement as a function of total citations in the prior 4 years, while controlling for the other variables in the model. The data points in the figure indicate that each additional prior traffic citation increases the odds of a subsequent accident by 10%.



Predicted probability of subsequent 4-year (1988-91) accident involvement as a function of total citations in the prior 4 years (1984-87).

The figure below displays the predicted probability and 95% confidence intervals for subsequent 4-year total accident involvement as a function of prior accident involvements, while controlling for other variables in the model. The data points in the figure indicate that each additional prior accident increases the odds of a subsequent accidents by 30%.



Predicted probability of subsequent 4-year (1988-91) accident involvement as a function of total accident involvement in the prior 4 years (1984-87).

Classification and prediction accuracy. The adequacy of the models was compared in terms of their accuracy for classification and prediction. The performance of the models is displayed in the table below. The numbers in the columns are counts of accidents actually incurred by drivers placed in designated risk groups by the models.

Number of Accidents During 1988-91 for Drivers Selected by Various Models as Being in Designated Risk Groups (*n* = 140,000)

	Drivers estimated by model to be in risk group				
Model	Highest 1,000	Highest 5,000	Highest 10,000	Highest 20,000	Highest 120,000
Prior neg-op points	453	1,963	3,619	6,539	25,174
Prior accidents	490	1,914	3,396	5,979	24,881
A1	555	2,237	3,949	7,017	25,740
A2	603	2,378	4,206	7,260	25,856
A3	567	2,261	4,029	7,075	25,791
B1	510	2,093	3,748	6,605	25,085
B2	543	2,143	3,904	6,906	25,268
B3	521	2,081	3,779	6,756	25,147
C1	535	2,205	3,916	6,996	25,769
C2	577	2,351	4,123	7,179	25,874
C3	535	2,220	3,984	7,017	25,816
D1	506	2,028	3,790	6,681	25,184
D2	560	2,217	3,965	6,940	25,313
D3	514	2,062	3,774	6,762	25,241
E1	597	2,346	4,121	7,227	25,884
E2	550	2,192	3,965	7,016	25,823
F	603	2,378	4,173	7,235	25,881
H	600	2,366	4,144	7,257	25,862

Note. Entries for prior neg-op points and prior accidents represent the number of drivers having the highest counts of these incidents during the prior 4-year (1984-87) period. For an explanation of model letter and number, see the previous table presented in this summary.

The results shown in the table support the following conclusions:

- A driver licensing agency can do better than to use either prior neg-op points or prior accidents alone in an attempt to identify drivers with a high probability of subsequent accident involvement.

- A driver's previous accident record is an important factor in estimating their future accident potential. For example, in estimating the worst 1,000 drivers, models A1, B1, and C1, which do not contain prior total accidents as predictors, perform more poorly than models using prior accidents as an explanatory variable.

- In assessing total accident "hits," models that employ prior responsible (at-fault) accidents as predictors (A3, B3, C3, and D3) do not perform as well as models using total accidents as predictors.

- Model combinations in A and C that use the demographic variables age, gender, and license class perform better than similar models that do not include these variables.

- Models that use total citations or the number of 0-, 1-, and 2-point citations perform better than models using the individual violation types.

- Models using separate counts of the number of TVS dismissals or of uncleared FTA violations (E1, E2, and H) improve prediction in identifying accident-involved drivers.

Predicting individual accident involvement. To illustrate the accuracy of the regression equations in predicting the future accident potential of individual drivers, 2x2 cross-classification tables were constructed displaying the relationship between each individual's predicted and actual accident-involvement frequency. The table below summarizes the results from these tables by displaying measures of classification accuracy for each model.

The following can be inferred from the table:

- Without exception, each model performs better than the one based on neg-op points alone ("Current Neg-Op"). For example, model D2, which uses prior total citations and accidents as predictors, correctly classified 26.51% of accident involved drivers, while the current neg-op point model accurately classified only 25.25% of the accident-involved drivers.

- While many of the same individual drivers are selected by the various models, the characteristics of any selected group of drivers are dependent on the model's predictor variables, which would determine the type of drivers that would be targeted for licensing action if the model were used for this purpose.

Percentage of Drivers Correctly Classified for Each Model

Model	Percent correctly classified	
	Accident-involved	Accident-free
A1	26.91	84.96
A2	27.60	85.10
A3	26.88	85.03
B1	25.95	84.73
B2	26.90	84.90
B3	26.20	84.81
C1	26.48	85.03
C2	27.38	85.06
C3	26.96	84.97
D1	25.59	85.00
D2	26.51	85.01
D3	25.43	85.15
E1	27.31	85.10
E2	26.94	84.97
F	27.57	85.09
H	27.58	85.09
Current Neg-Op	25.25	84.88

Note. A unique predicted accident probability cut-off was used to equalize the marginals for each model.

Conclusions

All of the models evaluated in this paper were consistent in demonstrating that increased probability of subsequent accident involvement is associated with increased prior citation and prior accident frequencies, being young, and being male. The results of these analyses are consistent with those of prior research using samples of California drivers.

The findings support the following conclusions:

- In an effort to identify high-risk drivers, a licensing agency can do better than to use either prior neg-op points or prior accidents alone.

- Prior accident involvements are an important factor in estimating future accident risk; however, models using culpable accidents do not perform as well as models using total accidents.

- Models that use demographic variables such as age, gender, and license class perform better than models that do not use these variables as predictors. The use of these demographic variables also results in improved accuracy of the models by reducing the number of false positives and false negatives. It should be noted, however, that the use of age and/or gender for selecting drivers for license control actions may not be legally or socially defensible.

- Model E1 yielded the greatest catch of high-risk drivers. This model used age, gender, license class, total citations, total accidents, and number of FTA violations as predictors. It was shown that among the 120,000 (out of 140,000) drivers with the worst predicted driving records, model E1 yielded 25,884 total accident hits during the next 4 years. Model F, which used age, gender, license class, and one parameter each for the number of 0-, 1-, and 2-point citations, yielded the second "richest" catch of high-risk drivers. It was shown that among the highest-risk 120,000 drivers, model F identified 25,881 total accident hits during the subsequent 4 years.

- Using the number of traffic violator school dismissals as an independent variable enhances performance of the accident prediction models. It has been well established in prior departmental reports that a TVS dismissal is significantly more predictive of future accidents than is an additional conviction.

- Comparisons of the different models confirm past findings that knowledge of individual violation types does not greatly improve the predictive capabilities of accident-prediction models.

- If the goal of driver record adjudication systems is to identify and apply sanctions to high-risk drivers in order to intervene before this risk is realized, then the results presented in this report support the current point-count strategy which attempts to optimize the identification of drivers having a high probability of subsequent accident involvement.

TABLE OF CONTENTS

APPENDICES

TABLE OF CONTENTS (Continued)

APPENDICES

TABLE OF CONTENTS (Continued)

APPENDICES (Continued)

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS (Continued)

LIST OF TABLES (Continued)

INTRODUCTION

Driver control systems that assign penalty points to various traffic law infractions, and establish a level of point accumulation at which a licensing action is taken, are used by almost all motor vehicle departments. However, in many cases, the number of penalty points assigned to each infraction type is basically a qualitative assessment without any empirical foundation. A large body of scientific literature indicates that it is difficult to accurately identify accident-prone drivers on the basis of their traffic accident and traffic conviction records (Peck, McBride, & Coppin, 1971; Peck & Kuan, 1983; Harano, McBride & Peck, 1973; Hauer, Persaud, Smiley & Duncan, 1991; Chen, Cooper, & Pinili, 1995). The primary objective of point systems is to identify and initiate driver improvement or license control actions against drivers who are the most likely to be involved in accidents. It is also commonly assumed that the existence of point systems serves as a general deterrent to negligent driving and the accumulation of numerous traffic citations.

In California, the negligent-operator (neg-op) point system operates as follows. Each conviction for a violation of the traffic law carries a certain number of neg-op points. For example, a cited driver gets one point charged against his driving record for a speeding violation and two points for a major violation such as driving under the influence of alcohol or drugs. Following conviction, these points are added to the driver's record. When the point count reaches specified levels, the driver is exposed to a "treatment." This treatment usually takes the form of a warning letter for the lowest specified neg-op point level and climaxes in the suspension or revocation of the driver license at the highest level of neg-op action.

Section 12810.5a of the California Vehicle Code (CVC) defines a *prima facie* negligent-operator as any Class C (passenger car) licensed driver "whose driving record shows a violation point count of four or more points in 12 months, six or more points in 24 months, or eight or more points in 36 months." Other sections of the CVC (13800 and 14250) grant the department discretionary authority to take a variety of license control actions, including license suspension, against drivers who meet the CVC's definition of a negligent-operator. Since the program is discretionary, the CVC (Section 13950) also requires that drivers be offered the opportunity for an administrative hearing pursuant to any actions proposed under the negligent-operator provisions. The point system for heavy-vehicle commercial drivers (Classes A and B) is different from that of Class C drivers as defined in CVC Section 12810.5b. For an overview of the findings and program improvements of California's negligent-operator treatment evaluation system from 1976 through 1995, the interested reader is referred to Peck and Healey (1995).

This paper will focus on identifying the negligent driver by predicting which drivers are most likely to have one or more accident involvements in a subsequent period of time on the basis of their biographical characteristics and their past record of traffic accidents and citations for traffic law violations.

A large body of research has established that statistically significant relationships exist between counts of traffic accident involvements and counts of traffic citations. Several of these studies have addressed in detail the estimation of a driver's future accident potential on the basis of prior driving record histories (e.g., Gebers, 1997; Peck, McBride, & Coppin, 1971).

More recently, Peck and Kuan (1983) identified two types of driver risk factors: (1) aggregate-level factors such as territory of residence and (2) person-centered variables such as prior record of convictions and accidents, age, gender, socioeconomic status, and driving exposure. They reported that driving record variables, driving exposure, and territory made unique contributions to accident prediction, and that person-centered driving record variables were substantially superior to aggregate-level variables in terms of their ability to predict future accident involvements. Peck and Kuan also found that driving record variables and exposure measures were approximately equally efficient as accident predictors.

An earlier study by McConnell and Hagen (1980) attempted to define and validate a method of identifying groups of high-risk drivers that would yield a more effective crash prediction model than would California's DMV neg-op point system in effect at that time. Based on a 3-year driver record, five high-risk groups were identified from a sample of over 200,000 licensed drivers. These high-risk groups included drivers with various combinations of major and minor traffic citations. For each of the five groups, a regression equation was developed to maximize the prediction of accident involvement in a future 3-year period. These equations were then cross-validated on independent samples that met the risk-group definitions. The drivers identified as being high-risk by this approach were compared to drivers identified as being high-risk using two alternative regression equations and the neg-op point approach. While the high-risk group approach proved more effective than the neg-op point approach in predicting future accidents, the regression equations using the sum of all convictions and all accidents were even more effective as crash-prediction models. Based on these findings, the authors recommended implementation of a regression equation model using weighted accident and citation data as the optimal system for selecting high-risk drivers for post-licensing control actions.

Gebers (1997) evaluated the relative importance and significance of the factors explaining the number of traffic accidents during a given time period. He employed a database consisting of the accident record and characteristics of individual drivers. The dependent variable was the number of accidents an individual had in the time period considered. The techniques evaluated consisted of ordinary least squares, weighted least squares, Poisson, negative binomial, linear probability, and logistic regression models. The results showed that the different regression methods produced almost identical results in terms of the relative importance and statistical significance of the independent variables.

The results presented by Gebers were similar to findings reported by Boyer, Dionne, and Vanasse (1990). These researchers evaluated traffic accidents by comparing the results estimated from both categorical and count-data models. Although the authors stressed the importance of selecting the appropriate model from quantitative predictors, it was shown that in all models the individual's past driving record is a relatively good predictor of future traffic accident risk.

In a recent study of driver accident risk in relation to the penalty point system in British Columbia, Chen, Cooper, and Pinili (1995) assessed the relative impact on future crash-involvement risk of a number of different infractions and also of accident history. These authors examined 1,998,347 British Columbia driver records. Logistic regression was used to identify drivers who were most likely to have one or more at-fault accident involvements in a post-period on the basis of their pre-period record of at-fault accident involvements and convictions. The results showed a consistent increase in post-period at-fault accidents per driver, with increasing pre-period numbers of both crashes and convictions. It was also found that prior at-fault accident involvements were a better predictor of future at-fault accidents than were prior traffic citations, and that up to 23% more high-risk drivers could be identified using prior culpable accidents than by traffic convictions alone. Of individual violation types, right-of-way violations such as failure-to-yield and disobeying a traffic signal were found, after accidents, to be the type of pre-period incidents most strongly associated with post-period at-fault traffic accident involvements. It should be noted that the authors did not include total or nonculpable accidents in their study, either as predictors or criterion variables. In addition, the authors did not demonstrate that the individual violation and accident types improved prediction beyond that achieved by use of their sums (i.e., total convictions and total culpable accidents).

Hauer, Persaud, Smiley, and Duncan (1991) examined person-centered variables to estimate the accident potential of Ontario drivers. Accident potential was defined as the expected number of accidents per unit of time. They compared 16 distinct models, for each of which parameters were estimated. The authors reported that the model that used detailed information on age, gender, individual violation types, and the count of at-fault and not-at-fault police-reported accidents was the most efficient one in explaining estimated accident potential. Hauer et al. (1991) concluded that to identify drivers with accident potential, one can do better than to use demerit points based on the perceived seriousness of convictions and that it is important to use the driver’s previous accident record. The authors also noted that models using only at-fault accidents perform worse than those using all accidents, and that models that include age and gender perform better than the corresponding models that do not.

The present paper will further explore issues of accident-risk modeling, building on the techniques presented in the cited works by Chen et al. (1995) and Hauer et al. (1991). The analyses in the following sections are designed to estimate (1) a driver’s accident risk (i.e., the expected probability of accident involvement during the subsequent accident criterion period) and (2) the accuracy of prediction (i.e., the rate of true positives and true negatives) using a variety of prediction models as discussed below. It is intended that the results from the analyses will be used to estimate the accident risk levels of the identifiable subgroups in the driver population and to assist in the ongoing refinement of California’s point system and other safety programs for selecting negligent and accident-prone drivers for treatment and driver control actions.

Before proceeding, it is important to review the caveat raised in Peck and Kuan (1983) in relation to the distinction between individual and group prediction when evaluating the efficacy of an accident prediction system. As stated in the above research efforts, prior accident record is predictive of subsequent accident record. However, as Peck and Kuan note, it is incorrect to conclude that the majority of accidents are caused by a small number of accident-prone drivers or that an individual's future accident involvement can be predicted with a high degree of precision from past accident involvement.

In fact, several studies have demonstrated that the majority of accidents in any time period involve drivers with so called "average" or "clean" prior driving records. This is essentially because there are many more drivers with average or good prior driving records than there are with bad ones, and also because accident involvement depends on many factors in addition to a driver's behavior at any point in time. The large random or stochastic component in accident causation (i.e., the variation in accident occurrence that is not systematically associated with measurable differences between people) makes it impossible to accurately predict which individuals will be involved in accidents (Peck, 1993; Peck et al., 1971). However, as demonstrated in the following sections, it is possible to predict the accident involvement frequencies for *groups* of drivers.

The fact that there is a large amount of randomness in determining accident occurrence does not imply that all drivers pose the same accident risk or that human error plays a negligible role in accident causation. One reason why driver negligence does not always cause a crash is that many of the accident-related human errors that all drivers sometimes commit (e.g., momentary lapse of attention) result in accidents only when there exists a complex set of conditions necessary for the accident to occur. On the other hand, even the very best driver can become involved in an accident because of the negligent actions of others. For these reasons, although certain types and groups of drivers have significantly higher accident rates than do others, the number of accident involvements for any single individual cannot be accurately predicted from the available data (Gebers, 1990; Peck & Kuan, 1983; Peck et al., 1971).

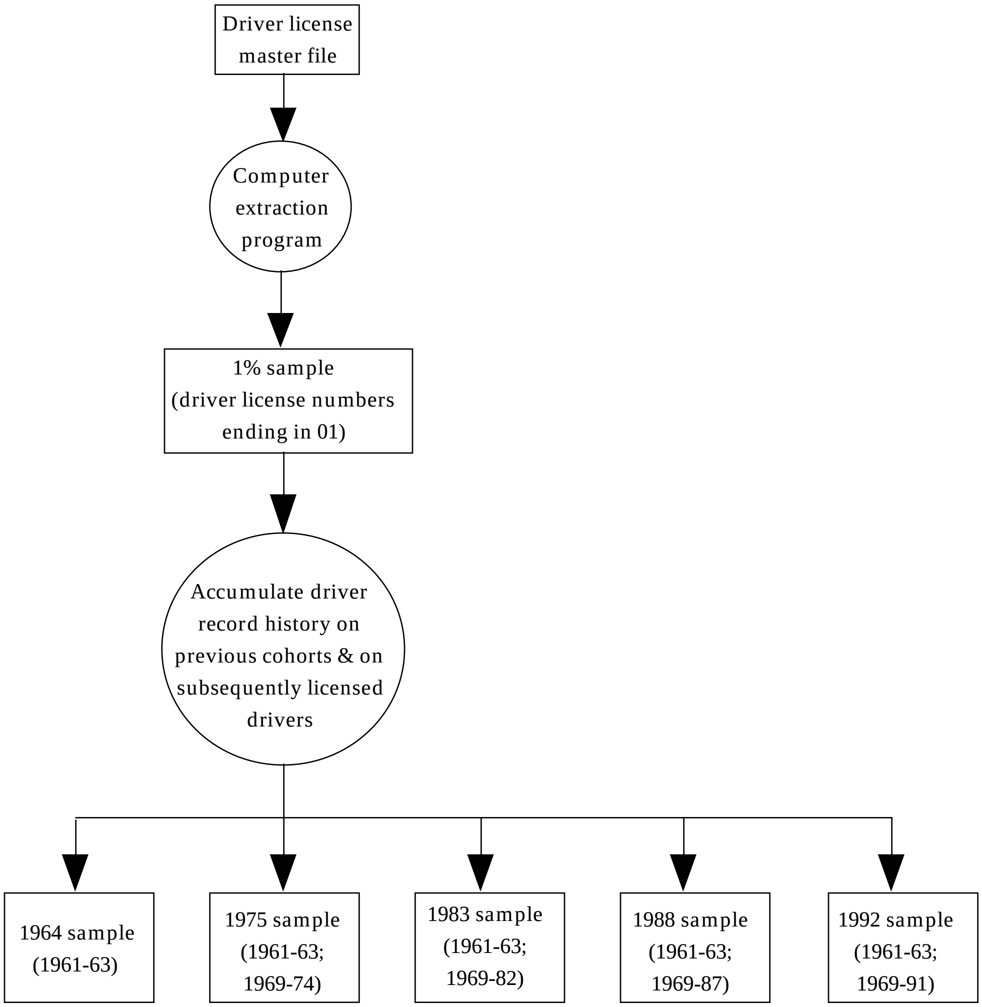
METHOD

Data

The California Department of Motor Vehicles (DMV) maintains an automated file containing driving records for over 20 million California drivers. The driver license (DL) number for each record consists of a letter prefix followed by a seven-digit numerical field. A 1% random sample of driver records, consisting of those with a DL number ending in 01, was extracted from the Department's master file on May 1, 1992 and merged into what is called the California Driver Record Study Database. These data served as the database for the present study.

Figure 1 summarizes the structure of the database from the California Driver Record Study used for the present analyses. As illustrated in the figure, a 1% random sample of the DL file has been extracted five times in the past, beginning in 1964. Driver record history data obtained from each extraction were merged, based on a matching of DL numbers, with data previously extracted for existing cohorts. In addition, all drivers in the sample who were not captured in the previous extractions entered the database and formed the basis for future tracking.

Data for the approximately 200,000 driver records extracted in 1992 include almost everything available on the DL file—demographic data, accidents and citations by type, physical and mental (P&M) codes, suspension/revocation (S/R) actions, and licensing variables such as class of license and driving restrictions. Driver record information stored on the California database covers the period 1961 through 1963 and 1969 through 1991. Data for 1964 through 1968 were purged from the DL file before they could be extracted and therefore are not in the database. The time period covered by an individual driver record is a function of when the driver was first licensed in California. To be included in the sample, individuals had to possess a valid California driver license at the time of the extraction. All drivers with a “deceased” indicator on their record or whose driver license had been expired for more than 6 months at the time of the extraction were excluded. The final study sample included approximately 140,000 drivers.


Driver license master file
Computer extraction program
1% sample (driver license numbers ending in 01)
Accumulate driver record history on previous cohorts & on subsequently licensed drivers
1964 sample (1961-63)
1975 sample (1961-63; 1969-74)
1983 sample (1961-63; 1969-82)
1988 sample (1961-63; 1969-87)
1992 sample (1961-63; 1969-91)

Note. The time periods in parentheses represent the years for which driver record histories are available in the database. Due to a purge of data from the department's DL master file, there are no data for 1964-68.

Figure 1. Process for creating the California Driver Record Study Database.

The data analyzed in the study were the following:

(1) Driver identification data—driver license number, county of residence, year and month of birth, and gender.

(2) Driver licensing data—month and year of license issuance, type of license issuance (e.g., new, renewal, duplicate, name change), test results, license class (e.g., noncommercial, commercial), type of driver license restrictions, year of expiration, months license expired, months license in force, and physical and mental disorders affecting driver performance.

(3) Driver record citation data—number of reported traffic citations by year for the 4-year period from 1984 through 1987. This includes summaries of one-point, two-point, and noncountable citations, as well as separate counts of citations of various types. The individual citation types consist of the following:

- Sign or signal (including traffic signs, signals, and markings)
- Roadway markings
- Lane placement
- Following too closely
- Unsafe passing and overtaking
- Right-of-way
- Turning
- Signaling
- Speed too fast
- Speed too slow
- Unsafe equipment
- Driver license restriction violations
- Driving without a license
- Driving under the influence of alcohol and/or drugs
- Reckless driving
- Driving with a suspended/revoked license
- Hit-and-run accident

The total-citations variable, which does not include uncleared failures to appear in court (FTAs), was based on traffic incidents and not the total number of citations for each incident. For example, if a driver received two citations on the same ticket (e.g., speeding and running a stop sign), this would be counted as only one incident for purposes of the total-citation variable. However, both of the citations would be counted separately under the appropriate citation-type variable.

(4) Traffic accident data—collected over the 8-year period ranging from 1984 through 1991. The data are presented for reported accidents only. California Vehicle Code Section 16000 requires the driver of each motor vehicle involved in an accident resulting in damage to the property of any one person in excess of $500, or in bodily injury or death of any person, to submit a written report to the Department of Motor Vehicles. Failure to file a report under the above conditions will result in the suspension of the driving privilege. Information was also collected on the responsibility or culpability of the accident as obtained from any accident report

filed by a law enforcement agency. An accident was categorized as an "at fault" accident if the official accident report found the involved driver to be the party most at fault or a party who contributed to the cause of the accident. The types of accidents investigated in this study consisted of total accidents and at-fault accidents.

(5) Negligent-operator points—in determining neg-op points in California, one point is entered on the driving record for each moving-violation conviction (e.g., speeding, unsafe turns), except those involving "major" offenses such as driving under the influence of alcohol/drugs, reckless driving, and hit-and-run. The latter convictions count two points each. An accident for which the driver is deemed at least partly responsible counts one point. As defined by CVC Section 12810.5, drivers with a Class C driver license are defined as neg-ops when their driver records contain four or more points in 1 year, six or more points in 2 years, or eight or more points in 3 years.

(6) Traffic Violator School (TVS) dismissals—traffic citations that were dismissed contingent upon completion of a state-certified TVS program as defined in CVC Section 42005. A citation that is dismissed conditional upon the offender's completion of TVS is not an actual conviction. In other words, TVS dismissals represent traffic citations that would not be counted if the analyses were limited to abstracts of traffic convictions.

(7) Uncleared FTAs—the number of uncleared failure-to-appear violations. These are violations under CVC Sections 40002 and 40508, which refer to citations for traffic violations in which the driver failed to keep his signed promise to appear in court.

Statistical Analyses

Multiple logistic regression was used to develop and assess a number of prediction models. Since the model produced by logistic regression is nonlinear, the equations used to predict the outcomes are slightly more complex than the more commonly used and familiar ordinary least squares regression equations. The interested reader is referred to texts such as Hosmer and Lemeshow (1989) and Tabachnick and Fidell (1996) for a detailed discussion of logistic regression analysis. The criterion variable is the estimated probability of one outcome (i.e., accident involvement), based on a nonlinear function of the best linear combination of predictors. With just two outcomes, the equation is

$$\hat{Y}_i = \frac{e^u}{1 + e^u}$$

where $\hat{Y}_i$ is the estimated probability that the i^{th} case (I = 1, ..., n) is in one of the outcome categories (i.e., Y = 1) and u is a product from the linear regression model:

$$u = A + B_1X_1 + B_2X_2 + \cdots + B_kX_k$$

with constant A, coefficients B_j, and predictors X_j for k predictors (j = 1, 2, ..., k).

The quantity u is the logit or natural log of the odds:

$$\ln\left(\frac{\hat{Y}}{1 - \hat{Y}}\right) = A + \Sigma B_j X_{ij}$$

That is, the linear regression term is the natural log of the probability of having one outcome (accident) divided by the probability of having the other outcome (no accident). The procedure for estimating coefficients is maximum likelihood, and the goal is to find the best linear combination of predictors to maximize the likelihood of obtaining the observed outcome frequencies.

Use of a logistic regression model allows for the computation of the odds of accident involvement for one group relative to those odds for another group; that is, an odds ratio. For example, if the odds for males (coded 1) and the odds for females (coded 0) were compared, an odds-ratio greater than 1 would indicate that males are a higher accident risk. A value of 1 would indicate that both sexes have equal odds of being in an accident. An odds-ratio of less than 1 would indicate that males are a lower accident risk.

Logistic regression is often used to fit and compare models. The simplest (and worst-fitting) model includes only the constant and no predictors. The most complex (and best-fitting) model includes the constant, all predictors, and, in some cases, interactions among the predictors. Goodness-of-fit tests are used to choose the model that does the best job of prediction with the fewest predictors. In the following sections, goodness-of-fit tests are applied to the estimated prediction models to compare:

- Models that use age, gender, and license class as predictor variables versus models that do not.

- Models that use at-fault or “responsible” accidents among the set of predictor variables versus models that use total accidents among the set of predictor variables.

- Models that have a separate parameter for each of the 17 individual citation types listed on page 8 as predictors versus models that have one common parameter for all citation types combined.

- Models that use the numbers of 0-, 1-, and 2-point citations as predictors versus models that do not.

- Models that use the number of uncleared FTA violations as a predictor versus models that do not.

- Models that use the number of TVS dismissals as a predictor versus models that do not.

- Models that use the number of neg-op points as a predictor versus models that do not.

Table 1 lists the models and variables that were evaluated. Analyses were conducted using SAS procedures FREQ, UNIVARIATE, and LOGISTIC (SAS Institute Inc., 1990a, 1990b).

Table 1

Predictor Variables Evaluated in Each Model

Mode l	Demographics (age, gender, license class)	Citations		Accidents		FTA	0, 1, & 2 points	Neg-op points	TVS dismissals
		17 types	Total	Responsible	Total				
A1	X	X							
A2	X	X			X				
A3	X	X		X					
B1		X							
B2		X			X				
B3		X		X					
C1	X		X						
C2	X		X		X				
C3	X		X	X					
D1			X						
D2			X		X				
D3			X	X					
E1	X		X		X	X			
E2	X		X	X		X			
F	X				X		X		
G								X	
H	X		X		X				X

Note. An X indicates the inclusion of the variables in the model.

For example, Table 1 indicates that model A1 estimates parameters for age, gender, license class, and the 17 individual violation types. In contrast, model A2 estimates parameters for age, gender, license class, each of the 17 individual violation types, and total accidents.

It should be noted that models C1 through E2 include TVS dismissals, normally associated with safety-related moving violations, in the count of total citations. However, model H includes TVS dismissals as a separate variable distinct from citations; so in this model TVS dismissals are excluded from the total-citation count.

The models were evaluated using a number of different techniques to determine (1) which model is preferable in identifying a driver's accident risk (i.e., expected probability of accident involvement during the criterion period), and (2) which model demonstrates the highest level of predictive accuracy as related to the rate of true positives and true negatives.

A series of 2 x 2 tables were constructed to classify observed outcomes vs. predicted outcomes for each model. Optimum prediction values, defined as the model equation value that results in the same distribution for predicted and observed outcomes and maximizes the phi-coefficient, were calculated for each model. (In this case, the phi-coefficient is the Pearson correlation coefficient between the two outcome categories.) Cutoff scores for each model were selected by generating predicted accident probabilities from the different equations and then iteratively retabulating the sample using different predicted probability scores until nearly equal marginal proportions were obtained. As a result of the equal marginal proportions, which gave equal weights to both types of errors and tended to maximize the overall accuracy of classification (i.e., the phi-coefficient), the cutoff score used for each model produced approximately equal numbers of false-negative and false-positive predictions.

RESULTS

Assessment of Citation Categories

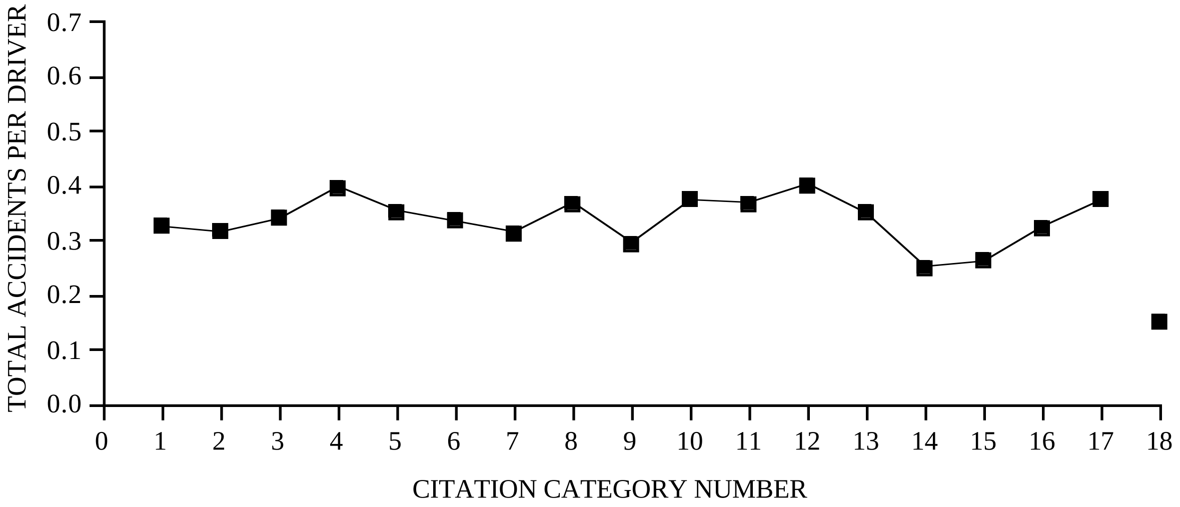
As stated above, a number of individual citation types were used in the development of several of the regression models. To assess the actuarial risk of drivers with different citation types, Table 2 displays the subsequent 4-year (1988-91) rate of total accidents by citation type for drivers with one or more citations in the prior 4 years (1984-87). These data are also displayed graphically in Figure 2.

Table 2 shows, for example, that 39,034 drivers were convicted of driving too fast during the period 1988-91. During the subsequent 4 years, these drivers accumulated 11,488 total accidents. This yields an average of 0.2943 accidents per driver. Similarly, the 6,729 drivers who were convicted of a turning violation in the prior 4 years have an average of 0.3142 accidents per driver in the subsequent 4 years.

Table 2

Subsequent 4-Year (1988-91) Total Accidents by Citation Type for Drivers with One or More Citations in the Prior 4 Years (1984-87)

Citation type	Number of drivers	Number of accidents	Mean
Sign or signal	16,795	5,484	0.3265
Roadway markings	1,123	356	0.3170
Lane placement	4,717	1,610	0.3413
Following too closely	1,651	655	0.3967
Unsafe passing & overtaking	1,509	532	0.3526
Right-of-way	2,514	849	0.3377
Turning	6,729	2,114	0.3142
Signaling	947	349	0.3685
Speed too fast	39,034	11,488	0.2943
Speed too slow	512	192	0.3750
Unsafe equipment	3,157	1,162	0.3681
DL restriction violations	253	102	0.4032
Driving without a license	6,001	2,109	0.3514
DUI	4,943	1,247	0.2523
Reckless driving	912	240	0.2632
Driving with S/R license	1,629	530	0.3254
Hit-and-run	240	90	0.3750
No citation	79,969	12,053	0.1507



1. Sign or signal
2. Roadway markings
3. Lane placement
4. Following too closely
5. Unsafe passing
6. Right-of-way
7. Turning
8. Signaling
9. Speed too fast
10. Speed too slow
11. Unsafe equipment
12. DL restriction violation
13. Driving without a license
14. DUI
15. Reckless driving
16. Driving with a S/R license
17. Hit-and-run
18. No citation

Note. Category #18 is a no-citation comparison group.

Figure 2. Subsequent 4-year (1988-91) total accidents for drivers with one or more citations in the prior 4 years (1984-87).

The future accident potential associated with each citation category is higher than the future accident potential of drivers with no reported citations (the last entry in the table) during the same prior 4-year period. This would imply that each violation category represents an accident potential and therefore should be retained as a candidate variable in the various regression models. Establishing accident potential is also necessary to enable a consistent comparison of the relative importance of each citation type in predicting accidents and in assessing risk as compared to “good” drivers with no prior citations. For example, the 16,795 drivers who are convicted of a sign or signal violation in the prior 4-year period have 2.17 (0.3265/0.1507) times as many accidents in the subsequent 4 years as drivers with no prior citations during the same prior 4-year period.

Assessment of Age Categories

It has been demonstrated that age is related to accident involvement (Gebers, Romanowicz, & McKenzie, 1990). Young drivers have consistently higher traffic accident rates than do older drivers. The data historically show that accident rates tend to decline through about age 69 and then increase.

In using age as a predictor of accident probabilities in the regression models, a couple of possibilities exist. One possibility is to use age as a continuous variable, with the option of making accident probability some polynomial function of age. (Prior work with these data indicated that accident potential tends to be a quadratic function of age.) A second possibility is to group ages into a number of distinct categories.

For the models presented in subsequent sections, age is divided into categories. It was decided to utilize age categories because the sample size is large (*n* = 140,000) and little information would be lost by aggregation. Additionally, the use of age categories avoided the need to build smoothing functions into the models, enhancing interpretability.

Figure 3 presents subsequent 4-year (1988-91) total accidents by age group. The figure shows that the accident rate is highest for the younger age groups. The accident rate declines until about age 69 and then begins to increase for the older age categories. In the subsequent analyses, the 45-49 year age group was selected as the reference, or comparison, category for the age variable.

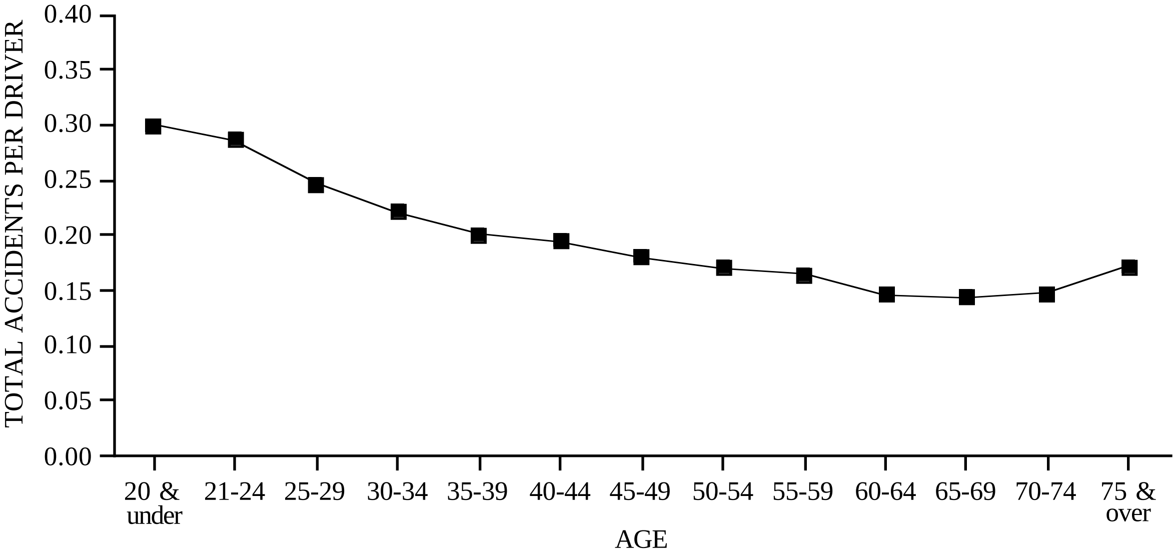


Figure 3. Subsequent 4-year (1988-91) total accidents by age for the 8-year sample.

Model Comparisons

In this section, parameter estimates are presented for the 17 logistic regression models defined in Table 1. As stated above, the comparisons involved models that do or do not use predictors (1) age, gender, and license class variables; (2) combinations of culpable and total accidents; (3) individual parameters for each of the 17 citation types; (4) uncleared FTAs; (5) 0-, 1-, and 2-point citations; (6) negligent-operator points; and (7) TVS dismissals.

The 17 logistic regression equations were evaluated by comparing the Akaike Information Criterion (AIC) statistic produced from each model. The AIC compares different models from the same data by adjusting the -2 Log Likelihood statistic for the number of terms in the model and for the number of observations in the sample.[1] The AIC value from a model consisting of only the intercept (A_I) and the AIC value from a model consisting of the intercept and variables (A_{I+V}) may be combined to form a statistic that compares fitted values of the models. The resulting value is interpreted as a proportion that measures how much better the model with intercept plus variables fits the data than does the intercept-only model. The relative AIC (AIC_{rel}) equation is as follows:

$$AIC_{rel} = \frac{(A_I - A_{I+V})}{A_I}$$

AIC_{rel} values obtained for the different models are presented in Table 3.

All of the relative AIC values are small. This implies that each model containing various combinations of the independent variables results in only a small (< 3%) improvement over the intercept-only model (i.e., a model that predicts all subjects to be at the mean probability of accident involvement). This suggests that the variable combinations from the models account for only a small part of accident potential and that other unknown or unidentified factors, as well as chance, account for nearly all of the variance in the accident outcome variable.

Table 3

Relative AIC Values for Different Models

	Model number within letter		
Model letter	1	2	3
A	0.022148	0.025452	0.022712
B	0.014673	0.018897	0.015526
C	0.021190	0.024415	0.021627
D	0.014659	0.018644	0.015308
E	0.024671	0.021938	––
F	0.025659	––	––
G	0.016787	––	––
H	0.025833	––	––

Note. See Table 1 for an explanation of model letters and numbers.

[1] AIC = -2 Log L + 2(k + s), where k = the number of ordered values for the response variable and s = the number of independent variables or covariates.

With this definition of the relative AIC values, one can make several inferences from Tables 1 and 3. The models that use prior total accidents as a predictor variable show increased relative AIC values over models that do not use total accidents as a predictor. Models that use prior culpable accidents as a predictor do not perform as well as models that use prior total accidents as a predictor. A comparison of the models in row A (in which the 17 individual violation types are used as predictors) to those in row C (in which only total citations is used as a predictor) shows only a small advantage of using the individual citation types. This finding is also evident in the comparisons of models in row B to the models in row D. However, comparisons of models in row A to the models in row B, and of the models in row C to the models in row D, convey the importance of using age, gender, and license class as estimators of accident probability. Models that use these demographic variables and various combinations of citations and accidents perform better than California's current neg-op system (model G), which uses a weighted combination of countable convictions and responsible accidents.[2]

For the sake of parsimony, only the two models with the highest relative AIC values will be discussed in detail. Table 3 shows that models F and H have the highest relative AIC values. Model F, which consists of age, gender, license class, the number of total accidents, and the number of 0-, 1-, and 2-point convictions (including dismissals) as predictors, had a relative AIC value of 0.025659. Model H, which consists of age, gender, license class, prior total citations, prior total accidents and (separately) the number of TVS dismissals as predictors, had a relative AIC value of 0.025833.

Table 4 summarizes the results of the nonconcurrent (prediction from earlier to later period) 8-year multiple logistic regression equation for predicting total accidents from model F. Table 5 presents the summary of the analogous logistic regression equation for model H. Appendix A presents summaries of the analogous regression equations for the remaining models. Asterisked odds ratios are significant at the .05 level.

Odds ratios greater than 1, if significant (asterisked), indicate enhanced risk. Odds ratios less than 1, if significant, indicate reduced risk. (Significance is shown when the interval between lower and upper confidence limits does not include 1.) Significant odds ratios showing increased risk, and the positive or negative direction of the regression coefficients in the two tables, indicate that increased probability of accident involvement is associated with:

- Being young.
- Being male.
- Holding a commercial driver's license.
- Increased prior citation frequency.
- Increased prior accident frequency.

[2] An exploratory analysis was performed only on violators, and it was found that even for this concentrated sample of deviant drivers, the total citations equation (R^2 = .0116) was comparable to the individual violation types equation (R^2 = .0121) in predicting the accident criterion.

Table 4

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model F (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	*p*	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.7728	0.0266	4429.57	.0001			
Age 20 & under	0.3599	0.0745	23.36	.0001	1.43*	1.24	1.66
Age 21-24	0.2995	0.0348	73.94	.0001	1.35*	1.26	1.44
Age 25-29	0.2211	0.0320	47.80	.0001	1.25*	1.17	1.33
Age 30-34	0.1665	0.0310	28.75	.0001	1.18*	1.11	1.26
Age 35-39	0.0957	0.0318	9.04	.0026	1.10*	1.03	1.17
Age 40-44	0.0694	0.0330	4.42	.0355	1.07*	1.01	1.14
Age 50-54	-0.0575	0.0380	2.29	.1301	0.94	0.88	1.02
Age 55-59	-0.0904	0.0394	5.27	.0216	0.91*	0.85	0.99
Age 60-64	-0.1545	0.0415	13.85	.0002	0.86*	0.79	0.93
Age 65-69	-0.1364	0.0434	9.88	.0017	0.87*	0.80	0.95
Age 70-74	-0.0939	0.0504	3.47	.0624	0.91	0.83	1.01
Age 75 & older	0.0352	0.0519	0.46	.4978	1.04	0.94	1.15
Gender	-0.2205	0.0152	209.97	.0001	0.80*	0.78	0.83
License class	0.5076	0.0337	226.68	.0001	1.66*	1.56	1.78
0-point citations	0.0394	0.0113	12.11	.0005	1.04*	1.02	1.06
1-point citations	0.1591	0.0059	727.53	.0001	1.17*	1.16	1.19
2-point citations	0.0957	0.0272	12.41	.0004	1.10*	1.04	1.16
Total accidents	0.2731	0.0131	437.87	.0001	1.31*	1.28	1.35

- 2 log likelihood for intercept only = 127,477.97
- 2 log likelihood for intercept and covariates = 124,170.97
χ^2 for covariates = 3,306.99, *p* = .0001

*Odds ratios are significant at the .05 level

For example, an examination of the signs of the regression coefficients and the 95% confidence limits for odds ratio values in Table 5 would lead to the following conclusions:

- Drivers aged 21-24 are 1.35 times as likely to be involved in a subsequent accident as are the comparison group of drivers aged 45-49.

- Drivers aged 65-69 are 0.88 times as likely (i.e., not as likely) to be involved in a subsequent accident as are drivers aged 45-49.

- Women are 0.80 times as likely (i.e., not as likely) to be involved in a subsequent accident as are men.

- Drivers with a commercial license are 1.67 times as likely to be involved in a subsequent accident as are drivers without a commercial license.

- Each additional prior traffic citation increases the odds of a subsequent accident by 10%.

- Each additional prior TVS dismissal increases the odds of a subsequent accident by 41%.

- Similarly, each additional prior traffic accident increases the odds of a subsequent accident by 30%.

In all cases, odds ratios for a particular variable are adjusted for the effect of other variables in the model.

Table 5

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model H (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.7757	0.0266	4450.85	.0001			
Age 20 & under	0.3357	0.0746	20.26	.0001	1.40*	1.21	1.62
Age 21-24	0.2961	0.0348	72.31	.0001	1.35*	1.26	1.44
Age 25-29	0.2102	0.0320	43.23	.0001	1.23*	1.16	1.31
Age 30-34	0.1584	0.0310	26.01	.0001	1.17*	1.10	1.25
Age 35-39	0.0911	0.0318	8.19	.0042	1.10*	1.03	1.17
Age 40-44	0.0688	0.0330	4.35	.0370	1.07	1.00	1.14
Age 50-54	-0.0559	0.0380	2.16	.1413	0.95	0.88	1.02
Age 55-59	-0.0858	0.0394	4.74	.0295	0.92*	0.85	0.99
Age 60-64	-0.1493	0.0415	12.94	.0003	0.86*	0.79	0.93
Age 65-69	-0.1301	0.0434	8.99	.0027	0.88*	0.81	0.96
Age 70-74	-0.0872	0.0504	3.00	.0834	0.92	0.83	1.01
Age 75 & older	0.0444	0.0519	0.73	.3922	1.05	0.94	1.16
Gender	-0.2187	0.0152	207.05	.0001	0.80*	0.78	0.83
License class	0.5101	0.0336	230.00	.0001	1.67*	1.56	1.78
Total citations	0.0932	0.0049	365.62	.0001	1.10*	1.09	1.11
TVS dismissals	0.3421	0.0170	402.65	.0001	1.41*	1.36	1.46
Total accidents	0.2636	0.0130	408.86	.0001	1.30*	1.27	1.34

- 2 log likelihood for intercept only = 127,477.97
- 2 log likelihood for intercept and covariates = 124,150.84
χ^2 for covariates = 3,327.13, p = .0001

*Odds ratios are significant at the .05 level

Figure 4 displays the predicted probability (mean) and 95% confidence intervals (upper confidence limits - UCL and lower confidence limits - LCL) of subsequent 4-year total accident involvement as a function of total citations in the prior 4 years, controlling for the other variables in the model.

Similarly, Figure 5 displays the predicted probability and 95% confidence intervals of subsequent 4-year total accident involvement as a function of prior accident involvements, controlling for the other variables in the model.

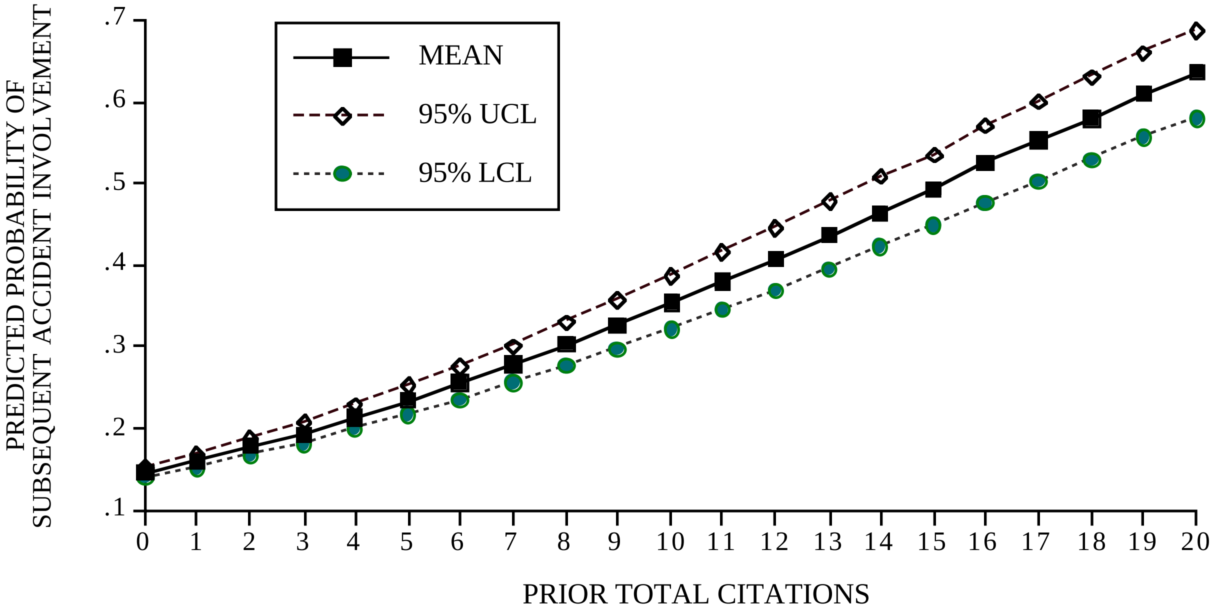


Figure 4. Predicted probability of subsequent 4-year (1988-91) accident involvement as a function of total citations in the prior 4 years (1984-87)

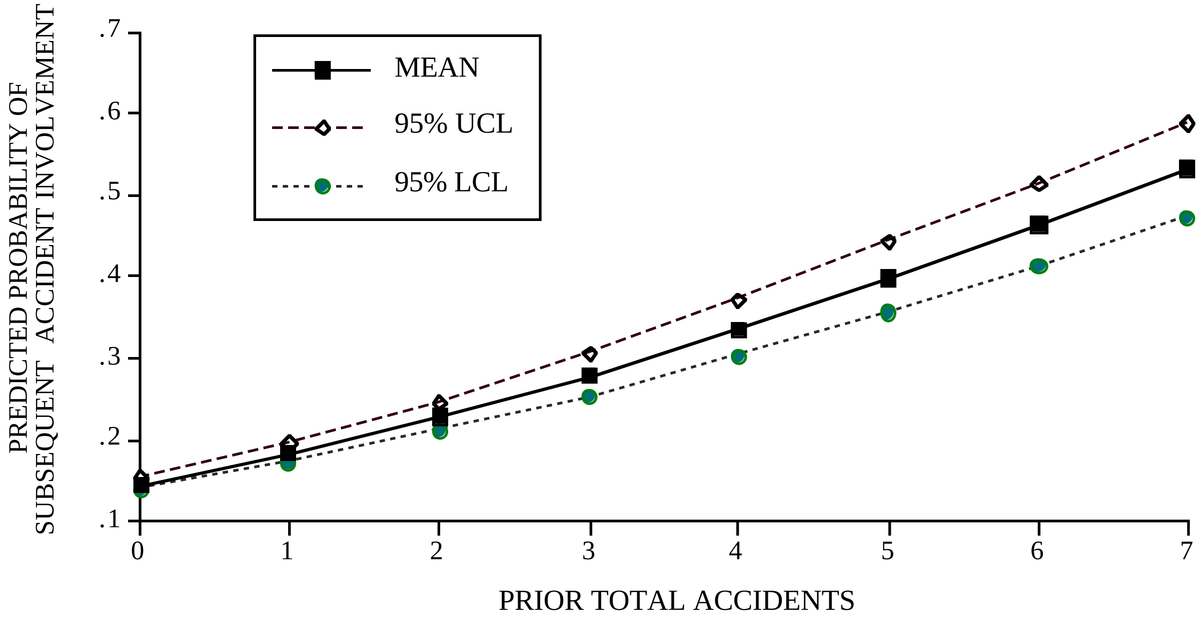


Figure 5. Predicted probability of subsequent 4-year (1988-91) accident involvement as a function of total accident involvement in the prior 4 years (1984-87).

Classification and Prediction Accuracy

In this section, measures of performance emanating from two strategies are presented to compare the adequacy of the different models in terms of classification and prediction accuracy. The first strategy identified the group of drivers with the most prior neg-op points during 1984-87, another group with the most prior total accidents during 1984-87, and 16 more groups estimated from the predicted scores in the different regression models as having the highest probability of accident involvement. Next, a count was made of the number of subsequent total accidents in which the drivers of each of these eightteen groups were involved during the subsequent 4-year (1988-91) period. The scheme or model that identified the most drivers who in 1988-91 accumulated the most accidents was deemed best. All models were evaluated at predicted probabilities of future accident involvement that produced equal numbers of high-risk drivers.

The second strategy focused on the accuracy of the models in predicting the subsequent accident status of the drivers (i.e., accident-involved versus accident-free). The false-negative and false-positive rates produced by the models were compared at a variety of predetermined cut-points in order to evaluate the respective sensitivity and specificity of the equations in predicting future accident involvement. The selected cutoff scores produced similar numbers of false-positive and false-negative predictions. Specificity is the proportion of no-event (i.e., accident-free) outcomes that were correctly predicted to be no-event. Sensitivity is defined as the proportion of the accident-involvement outcomes that were correctly predicted to be accident involved.

The performance of each of the 18 models (prior neg-op points, prior accidents, and the 16 regression models) is displayed in Table 6. The numbers in the columns are counts of accidents actually incurred by drivers placed in designated risk groups by the models. The results shown in Table 6 support the following conclusions:

- A driver licensing agency can do better than to use either prior neg-op points or prior accidents alone in an attempt to identify drivers with a high probability of subsequent accident involvement.

- A driver’s previous accident record is an important factor in estimating their future accident potential. For example, in estimating the worst 1,000 drivers, models A1, B1, and C1, which do not contain prior total accidents as predictors, perform more poorly than do models using prior accidents as an explanatory variable.

- In assessing total accident “hits,” models that employ prior responsible (at-fault) accidents as predictors (A3, B3, C3, and D3) do not perform as well as models using total accidents as predictors.

- Model combinations in A and C that use the demographic variables age, gender, and license class perform better than similar models that do not include these variables.

- Models that use total citations or the number of 0-, 1-, and 2-point citations perform better than models using the individual violation types.

- Models using separate counts of the number of TVS dismissals or of uncleared FTAs (E1, E2, H) add to the predictive value of identifying accident-involved drivers.

Table 6

Number of Accidents During 1988-91 for Drivers Selected by the Various Models as Being in Designated Risk Groups (*n* = 140,000)

	Drivers estimated by model to be in risk group				
Model	Highest 1,000	Highest 5,000	Highest 10,000	Highest 20,000	Highest 120,000
Prior neg-op points	453	1,963	3,619	6,539	25,174
Prior accidents	490	1,914	3,396	5,979	24,881
A1	555	2,237	3,949	7,017	25,740
A2	603	2,378	4,206	7,260	25,856
A3	567	2,261	4,029	7,075	25,791
B1	510	2,093	3,748	6,605	25,085
B2	543	2,143	3,904	6,906	25,268
B3	521	2,081	3,779	6,756	25,147
C1	535	2,205	3,916	6,996	25,769
C2	577	2,351	4,123	7,179	25,874
C3	535	2,220	3,984	7,017	25,816
D1	506	2,028	3,790	6,681	25,184
D2	560	2,217	3,965	6,940	25,313
D3	514	2,062	3,774	6,762	25,241
E1	597	2,346	4,121	7,227	25,884
E2	550	2,192	3,965	7,016	25,823
F	603	2,378	4,173	7,235	25,881
H	600	2,366	4,144	7,257	25,862

Note. Entries for prior neg-op points and prior accidents represent the number of drivers having the highest counts of these incidents during the prior 4-year (1984-87) period. For an explanation of model letter and number see Table 1.

These results are consistent with those reported above in relation to the relative AIC values.

The results in Table 6 indicate that the larger the pool of drivers (i.e., the lower the overall or average risk), the lower is the yield when identifying extremes of risk. This is to be expected. For example, among drivers selected by model A2, the 1,000 highest accident-risk drivers incurred a total of 603 accidents and thus had, on the average, approximately 0.603 accidents per driver over the subsequent 4-year period. This value is 3.02 times the 4-year average (0.200) for the total sample. Still considering A2, the 5,000 highest-risk drivers had 2,378 accidents, for an average of 0.476; the 10,000 highest-risk drivers had 4,206 accidents, for an average of 0.421.

The following section presents comparisons of the different models in terms of “hits,” “false alarms,” and “misses” in estimating individual accident involvement.

Predicting Individual Accident Involvement

Logistic regression equations can be conveniently used to predict, on the basis of an estimated probability score, the likelihood of a driver's accident involvement during a subsequent time period. Table 7 summarizes the possible classification outcomes from the logistic regression models.

Table 7

Crosstabulation of Predicted vs. Actual Outcomes

	Predicted outcome	
Actual outcome	Accident-involved	Accident-free
Accident-involved	*a* (true positive)	*b* (false negative)
Accident-free	*c* (false positive)	*d* (true negative)

As stated earlier, sensitivity is the proportion of event (here, accident-involvement) outcomes that were predicted correctly. Specificity is the proportion of no-event (here, no-accident) outcomes that were predicted correctly. The false-positive rate is the proportion of predicted accident outcomes which were wrong; no accident actually occurred. The false-negative rate is the proportion of predicted no-accident outcomes where the outcome was actually an accident.

With perfect prediction, all drivers would be counted in cells *a* and *d*, and no drivers would be counted in cells *b* and *c*. Drivers counted in cell *c* are false positives. These drivers are predicted to be accident-involved, but are actually accident-free. Drivers counted in cell *b* are false negatives. They are predicted to be accident-free, but are actually accident-involved. The predictive goal is to minimize the proportion of drivers in cells *b* and *c* and to make fewer errors than would be made in classifying drivers without the prediction equation. To be of any practical use, an equation must result in more classification accuracy than would be expected by chance alone.

To illustrate the accuracy of the regression equations in predicting the future accident potential of individual drivers, 2 x 2 cross-classification tables were constructed displaying the relationship between each individual's predicted and actual accident-involvement frequency.

Tables 8 and 9 present the fourfold 2 x 2 cross-classification tables for models A2 and C2, respectively. Recall that model A2 used demographic variables, total accidents, and the 17 violation types as predictors. Model C2 used the demographic variables, total accidents, and total citations as predictors. Each table used a different predicted-probability cutoff score for predicting whether a driver will have a future accident. The cutoff scores were selected by generating predicted accident probability scores from the different equations and then iteratively retabulating the sample using different cutoff scores until one was found that yielded nearly equal marginal proportions. As explained on page 13, the cutoff score used in each analysis also produced approximately equal numbers of false-negative and false-positive predictions, as would be expected from the equal marginal distributions. The use of equal marginals results in equal weights being assigned to both types of errors, and tends to maximize the overall accuracy of classifications, as represented by the phi-coefficient. Where one type of decision error has greater importance than does another, a different cutoff threshold can produce more optimal results. (The interested reader is referred to Peck and Kuan [1983] for the effects of using different cut-off thresholds on accident prediction models.)

Table 8 will be used here as an example of how to interpret the results. This table shows a statistically significant association ($p < .001$) between predicted and actual accident involvement in Model A2. The 23,843 drivers predicted to have accidents are almost 2 times as likely to be accident-involved as are the 115,642 drivers predicted to be accident-free (6,581 / [6,581 + 17,262] = 27.6% vs. 17,226 / [17,226 + 98,416] = 14.9%). However, the equation failed to correctly predict the majority of accident-involved drivers, as evidenced by the low true-positive rate of 27.6%. Although the false-negative rate of 14.9% appears low, this percentage of misclassification represents the majority of the 23,807 drivers (17.07% of the total sample) who were truly accident-involved.

The phi-coefficients and (accident) odds ratios shown at the bottom of Tables 8 and 9 are commonly used indices for quantifying the degree of association in contingency tables. As mentioned above, the phi-coefficient is simply the Pearson correlation coefficient between the actual and predicted accident-status categories. The odds ratio refers to the relative odds of being accident-involved for one predictive (accident) category relative to the other predictive (no accident) category. More specifically, the odds ratio is equal to $(P_a / P_c) / (P_b / P_d)$ or the cross-product ratio $P_a P_d / P_b P_c$, where the P_i represent the grand percentages in the respective cells (defined in Table 7).

Table 8

Actual Accident-Involvement Status by Predicted
Accident-Involvement Status for Model A2

	Predicted status		
Actual status	Accident-involved	Accident-free	Total
Accident-involved	6,581 (4.72%)	17,226 (12.35%)	23,807 (17.07%)
Accident-free	17,262 (12.38%)	98,416 (70.56%)	115,678 (82.93%)
Total	23,843 (17.09%)	115,642 (82.91%)	140,000 (100.00%)
Percent correctly classified	27.60%	85.10%	

Note. A predicted accident rate cutoff of 0.2107 was used to equalize marginals. The odds ratio is 2.18 and the phi-coefficient is .127.

Table 9

Actual Accident-Involvement Status by Predicted Accident-Involvement Status for Model C2

	Predicted status		
Actual status	Accident-involved	Accident-free	Total
Accident-involved	6,533 (4.68%)	17,274 (12.38%)	23,807 (17.07%)
Accident-free	17,329 (12.42%)	98,349 (70.51%)	115,678 (82.93%)
Total	23,862 (17.11%)	115,623 (82.89%)	140,000 (100.00%)
Percent correctly classified	27.38%	85.06%	

Note. A predicted accident rate cutoff of 0.2097 was used to equalize marginals. The odds ratio is 2.15 and the phi-coefficient is .124.

In Table 8, the probability of predicted accident-involved subjects actually having an accident as opposed to not actually having an accident (odds) is 0.3812 (4.72% / 12.38% or .2760 / .7240). Similarly, the odds of an accident for the predicted accident-free group are 0.1750 (12.35% / 70.56% or .1490 / .8510). The ratio of these two odds (i.e., the accident odds ratio) is 2.18. If the odds of having an accident did not vary as a function of predicted group, the odds ratio would be 1. This would imply that there is no difference between the prediction categories. Though this is not the case, the fact that the odds ratio and phi-coefficient are of modest size in Table 8 indicates that the predictive accuracy for individual drivers is not very good. This is demonstrated by the high false-positive rate and the fact that the equation misclassifies the majority of the accident-involved drivers.

Table 10 presents measures of classification accuracy for all of the models. A predicted accident probability cutoff score was used to equalize the marginals for each model. Every other model performs better than the one based on neg-op points alone. For example, model D2, which uses prior total citations and accidents as predictors, correctly classified 26.51% of accident involved drivers, while the current neg-op point model accurately classified only 25.25% of the accident-involved drivers.

Table 10

Percentage of Drivers Correctly Classified for Each Model

Model	Percent correctly classified		Phi	Odds ratio
	Accident-involved	Accident-free		
A1	26.91	84.96	.119	2.08
A2	27.60	85.10	.127	2.18
A3	26.88	85.03	.120	2.09
B1	25.95	84.73	.106	1.94
B2	26.90	84.90	.117	2.07
B3	26.20	84.81	.110	1.98
C1	26.48	85.03	.118	2.05
C2	27.38	85.06	.124	2.15
C3	26.96	84.97	.119	2.09
D1	25.59	85.00	.111	1.95
D2	26.51	85.01	.118	2.05
D3	25.43	85.15	.114	1.96
E1	27.31	85.10	.125	2.15
E2	26.94	84.97	.119	2.08
F	27.57	85.09	.127	2.17
H	27.58	85.09	.127	2.17
Current Neg-Op	25.25	84.88	.106	1.90

Note. A unique predicted accident probability cut-off was used to equalize the marginals for each model.

It is important to note that many of the same individual drivers are selected by the various models. This does not imply, however, that there are no differences in group membership across the different models. For example, Models B and D do not use gender, age, or license class as predictor variables. Drivers selected by these models will consist of fewer young males and commercial drivers than drivers selected by Model A or Model D, which do use these variables as predictors. Therefore, the characteristics of any selected group of drivers are dependent on the model's predictor variables, which will determine the type of drivers that would be targeted for licensing action were the model to be used for this purpose.

DISCUSSION

The goal of this paper was to assess the accuracy of predicting future accident risk using various combinations of demographic and prior driving record variables as predictors. The techniques presented were applied to California drivers and are a modification and extension of the methodology used by Smiley et al. (1989) and Hauer et al. (1991) in their studies of Ontario drivers, and by Chen et al. (1995) in their study of drivers in British Columbia.

All of the models were consistent in demonstrating that increased probability of subsequent accident involvement is associated with increased prior citation and prior accident frequencies, being young, and being male. The results of these analyses are also consistent with those of prior research using samples of California drivers (e.g., Gebers, 1997; Gebers & Peck, 1994; Peck & Gebers, 1992; Peck & Kuan, 1983).

The findings support the following conclusions:

- In an effort to identify high-risk drivers, a licensing agency can do better than to use either prior neg-op points or prior accidents alone. For example, the 120,000 drivers who in the prior 4-year period accumulated the most neg-op points had 25,174 accident involvements during the subsequent 4-year period. The 120,000 with the most accidents in the prior 4 years accumulated 24,881 accidents in the next 4 years. However, Model A2, which employs age, gender, license class, 17 individual citation types, and prior total accidents as predictors, yields a richer catch of high-risk drivers than does a count of either prior accidents or neg-op points. The highest-risk 120,000 drivers selected by this model incurred 25,856 accidents in the subsequent 4 years. This sum represents an increase of 3% over the 25,174 hits from using only neg-op points as the device for catching high-risk drivers, and an increase of almost 4% over the 24,881 hits from using prior accidents as the sole device for catching high-risk drivers.

- Prior accident involvements are an important factor in estimating future accident risk; however, not much is gained in differentiating between at-fault and total accidents as predictors.

- Models that use demographic variables such as age, gender, and license class perform better than do models that do not use these variables as predictors. The difference between such types of models becomes greater as one moves from the highest-risk 1,000 drivers selected by the models through the highest-risk 100,000 drivers selected. The use of these demographic variables also results in improved accuracy of the models by reducing the number of false positives and false negatives. It should be noted, however, that the use of age and/or gender as a device for selecting drivers for license control actions may not be legally or socially defensible.

- Model E1 yielded the greatest catch of high-risk drivers. This model used age, gender, license class, total citations, total accidents, and number of FTAs as predictors. Among the top 120,000 drivers with the worst predicted driving records, Model E1 yielded 25,884 total accident hits during the next 4 years. Model F, which used age, gender, license class, and one parameter each for the number of 0-, 1-, and 2-point citations, yielded the second “richest” catch of high-risk drivers. Among the highest-risk 120,000 drivers, model F identified 25,881 total accident hits during the subsequent 4 years.

- Using the number of traffic violator school dismissals as an independent variable enhances performance of the accident prediction models. It has been well established in prior research (e.g., Gebers, Tashima & Marsh, 1987; Peck & Gebers, 1991) that a TVS dismissal is significantly more predictive of future accidents than is an additional conviction.

- An additional model was produced in which the TVS and FTA variables were included in the same equation. Although not illustrated in this report, no appreciable difference in classification or prediction accuracy was evident beyond that reported for models in which the TVS and FTA variables were included separately.

- A comparison of the relative AIC values and measures of classification and predictive accuracy for the different models confirms past findings (e.g., McConnell & Hagen, 1980) that knowledge of individual violation types does not greatly improve the predictive capabilities of accident-prediction models above that of models using a count of the total number of citations.

If the goal of driver record adjudication systems is to identify and apply sanctions to high-risk drivers in order to intervene before this risk is realized, then the results presented in this report support refinements of the current point-count strategy to optimize the identification of drivers having a high probability of subsequent accident involvement. For example, driver licensing authorities may want to entertain incorporating driver age and the count of prior total traffic incidents (i.e., total citations regardless of point value and total accidents regardless of culpability) in the process through which high-risk drivers are defined and identified for possible treatment—by warning letter, hearing, probation, or the ultimate sanctions of license suspension and/or revocation.

REFERENCES

Boyer, M., Dionne, G., & Vanasse, C. (1990). *Econometric models of accident distributions*. University of Montreal: Center for Research on Transportation.

Chen, W., Cooper, P., & Pinili, M. (1995). Driver accident risk in relation to the penalty point system in British Columbia. *Journal of Safety Research*, *26*, 9-18.

Gebers, M. A. (1990). *Traffic conviction- and accident-record facts* (Report No. 127). Sacramento: California Department of Motor Vehicles.

Gebers, M. A. (1997). *Exploratory multivariable analyses of California driver record accident rates* (Report No. 166). Sacramento: California Department of Motor Vehicles.

Gebers, M. A., & Peck, R. C. (1994). *An inventory of California Driver accident risk factors* (Report No. 144). Sacramento: California Department of Motor Vehicles.

Gebers, M. A., Romanowicz, P. A., & McKenzie, D. M. (1993). *Teen and senior drivers* (Report No. 141). Sacramento: California Department of Motor Vehicles.

Gebers, M. A., Tashima, H. N., & Marsh, W. C. (1987). *Traffic violator school dismissals: The effects of citation masking on accident-risk assessment and on the volume of Department of Motor Vehicles' license control actions* (Report No. 113). Sacramento: California Department of Motor Vehicles.

Harano, R. M., McBride, R. S., & Peck. R. C. (1973). *The prediction of accident liability through biographical data and psychometric tests* (Report No. 39). Sacramento: California Department of Motor Vehicles.

Hauer, E., Persaud, B. N., Smiley, A., & Duncan, D. (1991). Estimating the accident potential of an Ontario driver. *Accident Analysis and Prevention*, *23*, 133-152.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York, NY: John Wiley & Sons.

McConnell, E. J., & Hagen, R. E. (1980). *Design and evaluation of a crash prediction strategy* (Report No. 76). Sacramento: California Department of Motor Vehicles.

Peck, R. C. (1993). 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, Drugs, and Driving*, *9*, 145-166.

Peck, R. C., & Gebers, M. A. (1991). *The traffic safety impact of TVS citation dismissals* (Report No. 133). Sacramento: California Department of Motor Vehicles.

Peck, R. C., & Gebers, M. A. (1992). *The California driver record study: A multiple regression analysis of driver record histories from 1969 through 1982*. Sacramento: California Department of Motor Vehicles.

Peck, R. C., & Healey, E. J. (1995). *California's negligent-operator treatment program evaluation system, 1976-95* (Report No. 155). Sacramento: California Department of Motor Vehicles.

Peck, R. C., & Kuan, J. (1983). A statistical model of individual accident risk prediction using driver record, territory and other biographical factors. *Accident Analysis and Prevention*, *15*, 371-393.

Peck, R. C., McBride, R. S., & Coppin, R. S. (1971). The distribution and prediction of driver accident frequencies. *Accident Analysis and Prevention*, *2*, 243-299.

SAS Institute Inc. (1990a). *SAS/STAT user's guide, version 6, volume 1* (4th Ed.). Cary, NC: Author.

SAS Institute Inc. (1990b). *SAS /STAT user's guide, version 6, volume 2* (4th Ed.). Cary, NC: Author.

Smiley, A., Persaud, B., Hauer, E., & Duncan, D. (1989). Accidents, convictions and demerit points: An Ontario driver records study. *Transportation Research Record*, *1238*, 53-64.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd Ed.). New York, NY: HarperCollins College Publishers.

APPENDIX

Table A1

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model A1 (n = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.7196	0.0265	4221.80	.0001			
Age 20 & under	0.4216	0.0742	32.32	.0001	1.52	1.32	1.76
Age 21-24	0.3392	0.0347	95.43	.0001	1.40	1.31	1.50
Age 25-29	0.2432	0.0319	58.19	.0001	1.28	1.20	1.36
Age 30-34	0.1798	0.0310	33.71	.0001	1.20	1.13	1.27
Age 35-39	0.1011	0.0318	10.13	.0015	1.11	1.04	1.18
Age 40-44	0.0699	0.0329	4.50	.0339	1.07	1.01	1.14
Age 50-54	-0.0578	0.0379	2.32	.1275	0.94	0.88	1.02
Age 55-59	-0.0943	0.0393	5.75	.0165	0.91	0.84	0.98
Age 60-64	-0.1592	0.0415	14.75	.0001	0.85	0.79	0.93
Age 65-69	-0.1469	0.0433	11.48	.0007	0.86	0.79	0.94
Age 70-74	-0.1041	0.0503	4.28	.0386	0.90	0.82	0.99
Age 75 & older	0.0337	0.0518	0.42	.5154	1.03	0.93	1.14
Gender	-0.2356	0.0152	241.49	.0001	0.79	0.77	0.81
License class	0.5758	0.0335	294.99	.0001	1.78	1.67	1.90
Sign or signals	0.2357	0.0168	196.66	.0001	1.27	1.23	1.31
Roadway markings	0.1187	0.0747	2.52	.1122	1.13	0.97	1.30
Lane placement	0.1820	0.0351	26.83	.0001	1.20	1.12	1.29
Following too close	0.3893	0.0567	47.10	.0001	1.48	1.32	1.65
Passing	0.2497	0.0611	16.69	.0001	1.28	1.14	1.45
Right-of-way	0.3133	0.0501	39.18	.0001	1.37	1.24	1.51
Turning	0.2669	0.0287	86.71	.0001	1.31	1.23	1.38
Signaling	0.1664	0.0830	4.01	.0451	1.18	1.00	1.39
Speed too fast	0.1468	0.0081	327.37	.0001	1.16	1.14	1.18
Speed too slow	0.1649	0.1130	2.13	.1445	1.18	0.94	1.47
Unsafe equipment	0.0580	0.0311	3.48	.0620	1.06	1.00	1.13
DL restrictions	-0.0886	0.1747	0.26	.6122	0.92	0.64	1.29
No DL	0.0188	0.0228	0.68	.4102	1.02	0.97	1.07
DUI	-0.0275	0.0289	0.90	.3415	0.97	0.92	1.03
Reckless driving	-0.1209	0.0853	2.01	.1564	0.89	0.75	1.05
Hit & run	0.2901	0.1641	3.13	.0770	1.34	0.96	1.83
14601	-0.0375	0.0322	1.36	.2443	0.96	0.90	1.03

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 124,592.54

χ^2 for covariates = 2,885.42, p = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A2

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model A2 (*n* = 140,000)

						Odds ratio 95% confidence limits	
Predictor variable	Regression coefficient	Standard error	Wald χ^2	*p*	Odds ratio	Lower	Upper
Intercept	-1.7664	0.0266	4403.50	.0001			
Age 20 & under	0.3732	0.0745	25.11	.0001	1.45	1.26	1.68
Age 21-24	0.3120	0.0348	80.28	.0001	1.37	1.28	1.46
Age 25-29	0.2296	0.0319	51.65	.0001	1.26	1.18	1.34
Age 30-34	0.1709	0.0310	30.33	.0001	1.19	1.12	1.26
Age 35-39	0.0979	0.0318	9.46	.0021	1.10	1.04	1.17
Age 40-44	0.0681	0.0330	4.26	.0390	1.07	1.00	1.14
Age 50-54	-0.0601	0.0380	2.51	.1134	0.94	0.87	1.01
Age 55-59	-0.0939	0.0394	5.69	.0171	0.91	0.84	0.98
Age 60-64	-0.1586	0.0415	14.60	.0001	0.85	0.79	0.93
Age 65-69	-0.1422	0.0434	10.74	.0010	0.87	0.80	0.94
Age 70-74	-0.0988	0.0504	3.85	.0498	0.91	0.82	1.00
Age 75 & older	0.0274	0.0519	0.28	.5972	1.03	0.93	1.14
Gender	-0.2270	0.0152	223.29	.0001	0.80	0.77	0.82
License class	0.5319	0.0337	248.45	.0001	1.70	1.59	1.82
Sign or signal	0.2153	0.0169	162.18	.0001	1.24	1.20	1.28
Roadway markings	0.1133	0.0749	2.29	.1303	1.12	0.97	1.30
Lane placement	0.1510	0.0353	18.32	.0001	1.16	1.09	1.25
Following too close	0.3623	0.0570	40.46	.0001	1.44	1.29	1.61
Unsafe passing	0.2285	0.0613	13.89	.002	1.26	1.11	1.42
Right-of-way	0.2307	0.0503	21.04	.0001	1.26	1.14	1.39
Turning	0.2419	0.0288	70.73	.0001	1.27	1.20	1.35
Signaling	0.1110	0.0833	1.77	.1829	1.12	0.95	1.32
Speed too fast	0.1318	0.0082	259.72	.0001	1.14	1.12	1.16
Speed too slow	0.1513	0.1134	1.78	.1821	1.16	0.93	1.45
Unsafe equipment	0.0391	0.0312	1.57	.2099	1.04	0.98	1.11
DL restrictions	-0.0948	0.1753	0.29	.5884	0.91	0.65	1.28
No DL	0.0196	0.0228	0.74	.3891	1.02	0.98	1.07
DUI	-0.0650	0.0292	4.98	.0257	0.94	0.89	0.99
Reckless driving	-0.1740	0.0858	4.11	.0425	0.84	0.71	0.99
Hit-and-run	0.1086	0.1651	0.43	.5108	1.12	0.81	1.54
14601	-0.0504	0.0324	2.42	.1198	0.95	0.89	1.01
Total accidents	0.2746	0.0131	440.99	.0001	1.32	1.28	1.35

- 2 log likelihood for intercept only = 127,477.97

- 2 log likelihood for intercept and covariates = 124,169.32

χ^2 for covariates = 3,308.65, *p* = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A3

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model A3 (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.7256	0.0265	4245.97	.0001			
Age 20 & under	0.4045	0.0743	29.67	.0001	1.50	1.30	1.73
Age 21-24	0.3273	0.0348	88.63	.0001	1.39	1.30	1.49
Age 25-29	0.2384	0.0319	55.86	.0001	1.27	1.19	1.35
Age 30-34	0.1780	0.0310	33.00	.0001	1.20	1.12	1.27
Age 35-39	0.0996	0.0318	9.83	.0017	1.11	1.04	1.18
Age 40-44	0.0701	0.0329	4.53	.0333	1.07	1.01	1.14
Age 50-54	-0.0573	0.0379	2.28	.1310	0.94	0.88	1.02
Age 55-59	-0.0941	0.0393	5.72	.0168	0.91	0.84	0.98
Age 60-64	-0.1606	0.0415	15.00	.0001	0.85	0.79	0.92
Age 65-69	-0.1475	0.0434	11.57	.0007	0.86	0.79	0.94
Age 70-74	-0.1059	0.0503	4.43	.0354	0.90	0.82	0.99
Age 75 & older	0.0257	0.0519	0.25	.6205	1.03	0.93	1.14
Gender	-0.2346	0.0152	239.41	.0001	0.79	0.77	0.82
License class	0.5600	0.0336	277.58	.0001	1.75	1.64	1.87
Sign or signals	0.2312	0.0168	188.67	.0001	1.26	1.22	1.30
Roadway markings	0.1134	0.0747	2.30	.1293	1.12	0.97	1.30
Lane placement	0.1699	0.0352	23.32	.0001	1.19	1.11	1.27
Following too close	0.3794	0.0568	44.67	.0001	1.46	1.31	1.63
Passing	0.2430	0.0612	15.78	.0001	1.28	1.13	1.44
Right-of-way	0.2631	0.0504	27.26	.0001	1.30	1.18	1.44
Turning	0.2625	0.0287	83.78	.0001	1.30	1.23	1.38
Signaling	0.1348	0.0831	2.63	.1049	1.14	0.97	1.35
Speed too fast	0.1431	0.0081	310.02	.0001	1.15	1.14	1.17
Speed too slow	0.1668	0.1129	2.18	.1396	1.18	0.95	1.47
Unsafe equipment	0.0512	0.0311	2.71	.0995	1.05	0.99	1.12
DL restrictions	-0.0913	0.1746	0.27	.6010	0.91	0.65	1.29
No DL	0.0169	0.0228	0.55	.4590	1.02	0.97	1.06
DUI	-0.0543	0.0292	3.46	.0629	0.95	0.90	1.00
Reckless driving	-0.1494	0.0855	3.05	.0806	0.86	0.73	1.02
Hit-and-run	0.1701	0.1651	1.06	.3028	1.19	0.86	1.64
14601	-0.0470	0.0324	2.11	.1466	0.95	0.90	1.02
Responsible accidents	0.2295	0.0262	76.61	.0001	1.26	1.20	1.32

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 124,518.61

χ^2 for covariates = 2,959.36, p = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A4

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model B1 (n = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits Lower	Odds ratio 95% confidence limits Upper
Intercept	-1.7552	0.0085	43020.33	.0001			
Sign or signals	0.2776	0.0167	275.83	.0001	1.32	1.28	1.36
Roadway markings	0.1825	0.0748	5.95	.0148	1.20	1.04	1.39
Lane placement	0.2316	0.0351	43.50	.0001	1.26	1.17	1.35
Following too close	0.4638	0.0568	66.76	.0001	1.59	1.42	1.78
Passing	0.2980	0.0612	23.72	.0001	1.35	1.20	1.52
Right-of-way	0.3282	0.0499	43.21	.0001	1.39	1.26	1.53
Turning	0.2942	0.0286	105.88	.0001	1.34	1.27	1.42
Signaling	0.2086	0.0832	6.28	.0122	1.23	1.05	1.45
Speed too fast	0.1967	0.0079	625.33	.0001	1.22	1.20	1.24
Speed too slow	0.2228	0.1132	3.87	.0490	1.25	1.00	1.56
Unsafe equipment	0.1554	0.0313	24.58	.0001	1.17	1.10	1.24
DL restrictions	-0.0173	0.1746	0.01	.9210	0.98	0.70	1.38
No DL	0.0566	0.0229	6.09	.0136	1.06	1.01	1.11
DUI	0.0449	0.0286	2.46	.1168	1.05	0.99	1.11
Reckless driving	-0.0524	0.0857	0.37	.5410	0.95	0.80	1.12
Hit-and-run	0.3706	0.1646	5.07	.0244	1.45	1.05	2.00
14601	-0.0370	0.0326	1.29	.2560	0.96	0.90	1.03

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 125,573.52

χ^2 for covariates = 1,904.44, p = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A5

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model B2 (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	*p*	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.8105	0.0089	41621.51	.0001			
Sign or signals	0.2518	0.0168	223.92	.0001	1.29	1.25	1.33
Roadway markings	0.1731	0.0749	5.34	.0209	1.19	1.03	1.38
Lane placement	0.1936	0.0353	30.12	.0001	1.21	1.13	1.30
Following too close	0.4281	0.0570	56.41	.0001	1.53	1.37	1.72
Passing	0.2712	0.0614	19.52	.0001	1.31	1.16	1.48
Right-of-way	0.2314	0.0503	21.19	.0001	1.26	1.14	1.39
Turning	0.2647	0.0287	85.08	.0001	1.30	1.23	1.38
Signaling	0.1431	0.0835	2.94	.0863	1.15	0.98	1.36
Speed too fast	0.1768	0.0080	494.65	.0001	1.19	1.18	1.21
Speed too slow	0.2042	0.1134	3.24	.0717	1.23	0.98	1.53
Unsafe equipment	0.1269	0.0313	16.43	.0001	1.14	1.07	1.21
DL restrictions	-0.0307	0.1754	0.03	.8609	0.97	0.69	1.37
No DL	0.0551	0.0229	5.80	.0161	1.06	1.01	1.11
DUI	-0.0011	0.0289	0.00	.9710	1.00	0.94	1.06
Reckless driving	-0.1145	0.0862	1.76	.1841	0.89	0.75	1.06
Hit-and-run	0.1590	0.1661	0.92	.3382	1.17	0.85	1.62
14601	-0.0508	0.0327	2.42	.1200	0.95	0.89	1.01
Total accidents	0.3093	0.0130	567.55	.0001	1.36	1.33	1.40

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 125,033.05

χ^2 for covariates = 2,444.91, *p* = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A6

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model B3 (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	*p*	Odds ratio	Odds ratio 95% confidence limits: Lower	Odds ratio 95% confidence limits: Upper
Intercept	-1.7652	0.0085	42823.12	.0001			
Sign or signals	0.2710	0.0167	262.25	.0001	1.31	1.27	1.36
Roadway markings	0.1753	0.0749	5.48	.0192	1.19	1.03	1.38
Lane placement	0.2157	0.0352	37.64	.0001	1.24	1.16	1.33
Following too close	0.4498	0.0568	62.70	.0001	1.57	1.40	1.75
Passing	0.2891	0.0612	22.31	.0001	1.34	1.18	1.51
Right-of-way	0.2648	0.0503	27.68	.0001	1.30	1.18	1.44
Turning	0.2888	0.0286	101.87	.0001	1.34	1.26	1.41
Signaling	0.1694	0.0833	4.14	.0419	1.19	1.01	1.40
Speed too fast	0.1913	0.0079	587.59	.0001	1.21	1.19	1.23
Speed too slow	0.2232	0.1130	3.90	.0482	1.25	1.00	1.56
Unsafe equipment	0.1442	0.0313	21.22	.0001	1.16	1.09	1.23
DL restrictions	-0.0244	0.1747	0.02	.8887	0.98	0.69	1.37
No DL	0.0534	0.0229	5.44	.0197	1.06	1.01	1.10
DUI	0.0112	0.0289	0.15	.6981	1.01	0.96	1.07
Reckless driving	-0.0879	0.0859	1.05	.3066	0.92	0.77	1.08
Hit-and-run	0.2200	0.1660	1.76	.1850	1.25	0.90	1.73
14601	-0.0482	0.0328	2.17	.1410	0.95	0.89	1.02
Responsible accidents	0.2807	0.0261	115.90	.0001	1.32	1.26	1.39

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 125,462.72

χ^2 for covariates = 2,015.25 , *p* = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A7

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model C1 (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	*p*	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.7179	0.0264	4227.82	.0001			
Age 20 & under	0.4282	0.0740	33.47	.0001	1.54	1.33	1.77
Age 21-24	0.3273	0.0347	88.90	.0001	1.39	1.30	1.49
Age 25-29	0.2241	0.0319	49.37	.0001	1.25	1.18	1.33
Age 30-34	0.1692	0.0310	29.87	.0001	1.18	1.12	1.26
Age 35-39	0.0957	0.0318	9.08	.0026	1.10	1.03	1.17
Age 40-44	0.0719	0.0329	4.77	.0290	1.08	1.01	1.15
Age 50-54	-0.0542	0.0379	2.04	.1529	0.95	0.88	1.02
Age 55-59	-0.0894	0.0393	5.17	.0230	0.92	0.85	0.99
Age 60-64	-0.1557	0.0414	14.13	.0002	0.86	0.79	0.93
Age 65-69	-0.1416	0.0433	10.69	.0011	0.87	0.80	0.95
Age 70-74	-0.1010	0.0503	4.03	.0447	0.90	0.82	1.00
Age 75 & older	0.0405	0.0518	0.61	.4343	1.04	0.94	1.15
Gender	-0.2223	0.0152	214.97	.0001	0.80	0.78	0.83
License class	0.5382	0.0334	259.41	.0001	1.71	1.60	1.83
Total citations	0.1340	0.0045	905.46	.0001	1.14	1.13	1.15

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 124,746.63

χ^2 for covariates = 2,731.33, *p* = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A8

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model C2 (n = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits Lower	Upper
Intercept	-1.7642	0.0266	4407.10	.0001			
Age 20 & under	0.3835	0.0743	26.63	.0001	1.47	1.27	1.70
Age 21-24	0.3028	0.0348	75.69	.0001	1.35	1.26	1.45
Age 25-29	0.2124	0.0320	44.17	.0001	1.24	1.16	1.32
Age 30-34	0.1614	0.0310	27.06	.0001	1.18	1.11	1.25
Age 35-39	0.0930	0.0318	8.54	.0035	1.10	1.03	1.17
Age 40-44	0.0698	0.0330	4.49	.0341	1.07	1.01	1.14
Age 50-54	-0.0568	0.0380	2.24	.1344	0.95	0.88	1.02
Age 55-59	-0.0897	0.0394	5.19	.0227	0.91	0.85	0.99
Age 60-64	-0.1561	0.0415	14.16	.0002	0.86	0.79	0.93
Age 65-69	-0.1382	0.0434	10.16	.0014	0.87	0.80	0.95
Age 70-74	-0.0965	0.0504	3.67	.0553	0.91	0.82	1.00
Age 75 & older	0.0324	0.0519	0.39	.5320	1.03	0.93	1.14
Gender	-0.2136	0.0152	197.68	.0001	0.81	0.78	0.83
License class	0.4999	0.0336	221.07	.0001	1.65	1.54	1.76
Total citations	0.1163	0.0046	652.52	.0001	1.12	1.11	1.13
Total accidents	0.2701	0.0130	430.55	.0001	1.31	1.28	1.34

- 2 log likelihood for intercept only = 127,477.97

- 2 log likelihood for intercept and covariates = 124,333.51

χ^2 for covariates = 3,144.454, p = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A9

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model C3 (n = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.7231	0.0264	4248.50	.0001			
Age 20 & under	0.4152	0.0741	31.40	.0001	1.52	1.31	1.75
Age 21-24	0.3178	0.0348	83.66	.0001	1.37	1.28	1.47
Age 25-29	0.2201	0.0319	47.59	.0001	1.25	1.17	1.33
Age 30-34	0.1678	0.0310	29.36	.0001	1.18	1.11	1.26
Age 35-39	0.0944	0.0318	8.84	.0029	1.10	1.03	1.17
Age 40-44	0.0720	0.0329	4.78	.0288	1.08	1.01	1.15
Age 50-54	-0.0538	0.0379	2.02	.1555	0.95	0.88	1.02
Age 55-59	-0.0894	0.0393	5.18	.0229	0.91	0.85	0.99
Age 60-64	-0.1574	0.0414	14.43	.0001	0.85	0.79	0.93
Age 65-69	-0.1428	0.0433	10.86	.0010	0.87	0.80	0.94
Age 70-74	-0.1029	0.0503	4.19	.0407	0.90	0.82	1.00
Age 75 & older	0.0325	0.0518	0.39	.5299	1.03	0.93	1.14
Gender	-0.2209	0.0152	212.05	.0001	0.80	0.78	0.83
License class	0.5264	0.0335	247.28	.0001	1.69	1.59	1.81
Total citations	0.1282	0.0045	802.62	.0001	1.14	1.13	1.15
Responsible accidents	0.1995	0.0258	59.59	.0001	1.22	1.16	1.28

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 124,688.94

χ^2 for covariates = 2,789.02, p = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A10

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model D1 (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.7576	0.0084	43373.17	.0001			
Total citations	0.1809	0.0041	1955.25	.0001	1.20	1.19	1.21

-2 log likelihood for intercept only = 127,477.97
-2 log likelihood for intercept and covariates = 125,607.27
χ^2 for covariates = 1,870.70, p = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A11

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model D2 (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.8104	0.0088	41971.22	.0001			
Total citations	0.1584	0.0042	1409.72	.0001	1.17	1.16	1.18
Total accidents	0.2993	0.0129	535.06	.0001	1.35	1.32	1.38

-2 log likelihood for intercept only = 127,477.97
-2 log likelihood for intercept and covariates = 125,097.18
χ^2 for covariates = 2,380.78, p = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A12

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model D3 (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.7656	0.0085	43187.54	.0001			
Total citations	0.1730	0.0042	1709.97	.0001	1.19	1.18	1.20
Responsible accidents	0.2416	0.0257	88.21	.0001	1.27	1.21	1.34

-2 log likelihood for intercept only = 127,477.97
-2 log likelihood for intercept and covariates = 125,522.54
χ^2 for covariates = 1,955.43, p = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A13

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model E1 (n = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	p	Odds ratio	Odds ratio 95% confidence limits	
						Lower	Upper
Intercept	-1.7737	0.0266	4433.49	.0001			
Age 20 & under	0.3735	0.0744	25.19	.0001	1.45	1.26	1.68
Age 21-24	0.3050	0.0348	76.79	.0001	1.36	1.27	1.45
Age 25-29	0.2166	0.0320	45.93	.0001	1.24	1.17	1.32
Age 30-34	0.1630	0.0310	27.59	.0001	1.18	1.11	1.25
Age 35-39	0.0939	0.0318	8.70	.0032	1.10	1.03	1.17
Age 40-44	0.0696	0.0330	4.46	.0347	1.07	1.01	1.14
Age 50-54	-0.0569	0.0380	2.24	.1341	0.95	0.88	1.02
Age 55-59	-0.0892	0.0394	5.13	.0235	0.92	0.85	0.99
Age 60-64	-0.1543	0.0415	13.82	.0002	0.86	0.79	0.93
Age 65-69	-0.1351	0.0434	9.71	.0018	0.87	0.80	0.95
Age 70-74	-0.0922	0.0504	3.35	.0672	0.91	0.83	1.01
Age 75 & older	0.0373	0.0519	0.52	.4722	1.04	0.94	1.15
Gender	-0.2105	0.0152	191.70	.0001	0.81	0.79	0.84
License class	0.4985	0.0336	219.63	.0001	1.65	1.54	1.76
Total citations	0.1342	0.0055	597.52	.0001	1.14	1.13	1.16
FTA	0.0572	0.0113	25.54	.0001	1.06	1.04	1.08
Total accidents	0.2673	0.0130	421.09	.0001	1.31	1.27	1.34

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 124,298.9

χ^2 for covariates = 3,179.063, p = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A14

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model E2 (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	*p*	Odds ratio	Odds ratio 95% confidence limits Lower	Upper
Intercept	-1.7341	0.0265	4279.66	.0001			
Age 20 & under	0.4040	0.0742	29.64	.0001	1.50	1.30	1.73
Age 21-24	0.3202	0.0348	84.89	.0001	1.38	1.29	1.47
Age 25-29	0.2247	0.0319	49.57	.0001	1.25	1.18	1.33
Age 30-34	0.1695	0.0310	29.93	.0001	1.19	1.12	1.26
Age 35-39	0.0954	0.0318	9.02	.0027	1.10	1.03	1.17
Age 40-44	0.0717	0.0329	4.74	.0295	1.07	1.01	1.15
Age 50-54	-0.0540	0.0379	2.03	.1545	0.95	0.88	1.02
Age 55-59	-0.0889	0.0393	5.11	.0238	0.92	0.85	0.99
Age 60-64	-0.1554	0.0414	14.05	.0002	0.86	0.79	0.93
Age 65-69	-0.1393	0.0433	10.33	.0013	0.87	0.80	0.95
Age 70-74	-0.0981	0.0503	3.80	.0512	0.91	0.82	1.00
Age 75 & older	0.038	0.0518	0.54	.4639	1.04	0.94	1.15
Gender	-0.2174	0.0152	205.10	.0001	0.81	0.78	0.83
License class	0.5247	0.0335	245.37	.0001	1.69	1.58	1.81
Total citations	0.1477	0.0055	733.76	.0001	1.16	1.15	1.17
FTA	0.0635	0.0113	31.62	.0001	1.07	1.04	1.09
Responsible accidents	0.1961	0.0258	57.61	.0001	1.22	1.16	1.28

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 124,647.36

χ^2 for covariates = 2,830.606, *p* = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.

Table A15

Summary of Nonconcurrent 8-Year (1984-87; 1988-91) Multiple Logistic Regression Equation for Predicting Accident Involvement from Model G (*n* = 140,000)

Predictor variable	Regression coefficient	Standard error	Wald χ^2	*p*	Odds ratio	Odds ratio 95% confidence limits Lower	Upper
Intercept	-1.7577	0.0085	42826.66	.0001			
Neg-op points	0.1919	0.0044	1881.78	.0001	1.21	1.20	1.22

-2 log likelihood for intercept only = 127,477.97

-2 log likelihood for intercept and covariates = 125,704.29

χ^2 for covariates = 1,773.678, *p* = .0001

Note. The odds ratios are statistically significant if their confidence intervals do not include 1.