STATE OF CALIFORNIA
DEPARTMENT OF MOTOR VEHICLES

USING TRAFFIC CONVICTION CORRELATES TO IDENTIFY HIGH ACCIDENT-RISK DRIVERS

JUNE 2000

Licensing Operations Division
California Department of Motor Vehicles

Authors: Michael A. Gebers and
Raymond C. Peck
Research and Development Branch

RSS-00-187

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2000	

4. TITLE AND SUBTITLE
Using Traffic Conviction Correlates to Identify High Accident-Risk Drivers

5. FUNDING NUMBERS

6. AUTHOR(S)
Michael A. Gebers and Raymond C. Peck

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-00-187

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)

This study further explored previous research involving the viability of predicting accidents from equations constructed to predict convictions for the general driving population. Models that better identify drivers at increased risk of future accident involvement will increase the number of accidents prevented through post license control actions.

Although the results do not support the hypothesis that equations keyed to citations do as well as or better than equations keyed to accidents in predicting subsequent accidents, the results suggest that identification of future accident-involved drivers can be improved by either of two approaches. The first is to construct equations based on a combination of prior accidents and citations. California's neg-op system basically reflects such an approach since points are allocated to traffic convictions and culpable accidents. The second alternative is more elaborate, involving a truly multivariate approach in which the prediction equation consists of a two-variable vector of subsequent citations and accidents. The canonical correlation analysis performed for this study resulted in two orthogonal canonical functions or roots: A driving-incident function consisting of primarily citations and secondarily accidents and an almost exclusively accident function.

The results reported in this study indicate that subsequent driving record can be predicted from prior driving record for groups of individuals; however, the error rates at the individual level are inherently large. The models derived from the canonical analysis, while superior to the simpler models, would be very difficult to implement operationally. The most obvious problem relates to its complexity. Canonical correlation is difficult to comprehend. Another problem is that the equations contain a number of variables (e.g., age and gender) that would not be legally defensible in taking license control actions. This problem could be rectified, with some sacrifice in predictive power, by deleting the unacceptable variables. In addition, use of variables such as age and gender might be permissible for triggering educational advisory interventions.

14. SUBJECT TERMS
Motor vehicle accidents, traffic safety, accident proneness, accident rates, accident risk, convictions, high-risk drivers, multivariate analysis, regression analysis, regression models

15. NUMBER OF PAGES
26

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	None

NSN 7540-01-280-5500

Standard Form 298 (2-89)
Prescribed by ANSI Std. Z39-18
298-102

PREFACE

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

ACKNOWLEDGMENTS

This report was prepared by the Research and Development Branch of the California Department of Motor Vehicles. The major part of this report was conducted under the supervision of Robert A. Hagge, Research Manager. The authors wish to acknowledge the contribution of Dr. Mary Janke, retired Acting Research Chief, who reviewed and edited the latter drafts of this report.

Appreciation also goes to Douglas Luong, Management Services Technician, for preparing the report drafts, and Debbie McKenzie, Associate Government Program Analyst, for proofreading the report.

EXECUTIVE SUMMARY

Background

- One of the primary objectives of the California DMV is to protect the public from drivers who represent unacceptably high accident risks.

- Optimum fulfillment of this objective requires the development and execution of strategies for identifying high risk drivers.

- One such system is the negligent operator point system as defined in section 12810 of the California Vehicle Code. This statute assigns points to moving violations and accidents and authorizes the department to take driver control actions against drivers who meet the prima face definition of "negligent operator."

- The California DMV has conducted a number of research studies aimed at improving the validity of point systems in identifying or predicting drivers with a relatively high likelihood of being involved in future accidents.

- Equations designed to predict future accident involvement from involvement in prior accidents and other predictor variables have had low accuracy.

- In a study predicting accidents and convictions for a group of negligent operators, Harano (1975) found that the equation developed to predict convictions actually predicted subsequent accidents almost as well as the equation developed to predict accidents.

- Marsh and Hubert (1974) found in a study of negligent operators attending either group or individual hearings that predicted convictions produced higher cross-

validation coefficients with accidents than did the equation developed for predicting total accident involvement.

- Peck, McBride, and Coppin (1971) concluded that probably the most important factors contributing to the superiority of convictions as a predictor of total accidents are their greater frequency of occurrence as compared to accidents and the inclusion of accident-related elements in citation frequency. The authors noted that this combination of increased stability and the intrinsic overlapping of behavioral elements between accident and violation behavior makes citation frequency a better predictor of accidents.

- The present study further explored the viability of predicting accidents from equations constructed to predict convictions for the general driving population. Equations or models that better identify drivers at increased risk of future accident involvement would increase the number of accidents prevented through post license control actions.

Research Methods

- Data for the analyses were obtained from the driving records of a 1% random sample of licensed California drivers. Information was collected on: Driver age; gender; presence of a physical or mental condition on record; presence of driver license restrictions on record; number of citations during 1986-88; number of citations during 1986-88; number of accidents during 1986-88; number of accidents during 1989-91; and territorial variables within ZIP-Code of residence.

- Multiple linear regression analysis and canonical correlation analysis techniques were used to identify the combination of variables providing the most accurate prediction of the total accident criterion measure.

- A construct sample and a cross-validation sample were created for both the multiple regression and canonical correlation analyses. Regression coefficients derived from the construct-sample equations were applied to the cross-validation sample to test their validity. Validity coefficients were computed by correlating actual and predicted criterion values for the cross-validation sample.

Results

- The results of the analyses are consistent with those of prior traffic safety research, with all of the models indicating that increased accident involvement was associated with the following:

  – Increased prior citations
  – Increased prior accidents
  – Having a commercial driver license
  – Being young
  – Being male
  – Having a physical or mental condition on record
  – Having a driver license restriction on record

- The results do not support the hypothesis that equations keyed to citations do as well as or better than equations keyed to accidents in predicting subsequent accident involvement. For example, the multiple regression equation keyed to accidents resulted in correctly predicting or identifying 22.9% of the accident-involved drivers, while the equation keyed to citations identified only 20.9% of these drivers. However, as noted below, an approach (canonical correlation) which considered subsequent accidents and citations correlates simultaneously produced improved prediction.

- The canonical correlation technique was substantially superior to multiple regression analysis in predicting accident-involved drivers. This was evidenced by the phi coefficient of .156 for the canonical variates of accidents and citations compared to the phi coefficients of .109 and .102 for the accident and citation mediated multiple regression equations, respectively.

- The ability of the canonical equation to correctly predict accident involvement was also higher than that of either the citation or accident equation alone, as evidenced by the 26.2% accuracy rate in identifying accident-involved drivers. This compares to the 22.8% accuracy rate produced by the accident equation, representing a 14.8% increase in predictive accuracy.

Conclusions

- The results indicate that the relative risk levels for groups of individuals can be predicted from prior driving records. However, the ability to predict which individuals will be involved in accidents is extremely low.

- The results presented in this paper contradict earlier findings of Harano (1975) and Marsh and Hubert (1974). Failure to replicate the findings of these earlier studies is probably due to the differences in the study populations. The present study utilized a random sample of all California drivers. The earlier studies were restricted to a sample of negligent operators.

- The identification of future accident-involved drivers can be improved by either of two approaches. The first is to construct equations based on a combination of prior accidents and citations. California's negligent-operator point system reflects such a strategy since points are allocated to traffic convictions and responsible accidents. The second alternative is more elaborate and involves a truly multivariate approach in which the prediction equation consists of a two-variable vector of subsequent citations and accidents. The canonical variate score produced from the driving-incident functions consisting of primarily citations and secondarily accidents resulted in an accident "hit rate" significantly higher than the other equations evaluated.

- While superior to the simpler models, the model produced from the canonical analysis would be very difficult to implement operationally due to its complexity. Canonical correlation analysis is difficult to comprehend. The task of explaining a canonical-based point system to administrators, legislators, and the public would be very difficult.

- A problem with both the canonical and multiple regression models is that, in contrast to the department's neg-op point system, they contain a number of variables such as age and gender that would not be legally defensible in taking license control actions (although it might be permissible for triggering educational or advisory interventions). The problem could be resolved with some sacrifice in predictive power by deleting these variables.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Equations designed to predict accidents from prior accident involvements and other variables have resulted in low multiple *R*s and low accuracy in classifying drivers involved in 0 versus 1 or more accidents (usually less than 30% correctly classified).

Harano (1975) developed a model for predicting accidents and convictions for 2,337 negligent operators subsequent to their being sent a notice to attend a group driver improvement session. Driver record and criminal record data were obtained for all subjects. Subjects who attended the sessions completed questionnaires and psychological tests.

A novel finding of the Harano study was that the equation developed to predict convictions actually predicted future accidents in the cross-validation sample about as well as the equation developed to predict accidents. In fact, the cut-off score for predicted convictions was more accurate in classifying accident and non-accident subjects than was the cut-off score for the accident prediction equation.

Marsh and Hubert (1974) conducted a study in which 13,594 negligent operators, attending either group meetings or individual hearings, each filled out two questionnaires. Equations predicting post-contact accidents and convictions were developed by using stepwise multiple regression analyses on half of the sample. The authors reported that predicted convictions consistently produced higher cross-validation coefficients with accidents (i.e., better prediction in the other half of the sample) than did the accident prediction equation.

However, several qualifications must be addressed when considering the Harano and the Marsh and Hubert results. For example, in the Harano study, the accident analyses were primarily limited to police-reported accidents rather than to all accidents. The author reported that preliminary analyses of his data indicated police-reported accidents to be more reliable than total accidents. This was not the case in the Marsh and Hubert study, which utilized the total accident criterion. Another qualification is that both studies were based on negligent drivers. The fact that the negligent-operator population is a relatively homogenous group on factors such as gender, age, and prior driving records compared to the driving population as a whole operates to restrict variability and consequently may have attenuated the predictability of subsequent accidents and convictions from prior driver record measures.

One reason for the superiority of traffic convictions as a criterion measure is their greater reliability. Using a Poisson fitting technique attributed to Newbold (1927) and Cobb (1940), Peck, McBride, and Coppin (1971) reported a negative binomial coefficient of .37 for a distribution of 3-year accident rates, which represents the maximum theoretical correlation that can be obtained from an "infallible" or error-free set of predictor variables, given the accident distribution in their study. In contrast, 3-year traffic conviction rates yielded a much higher coefficient of .52. Test-retest reliabilities, in which the number of driving record entries in a prior period are correlated with subsequent entries, are also much higher for convictions than for accidents.

The comparative superiority of prior traffic citation frequency over prior accident frequency as a predictor of subsequent accident involvements was extensively addressed in papers authored by Burg (1967, 1975) and Peck, McBride, and Coppin (1971). The latter authors offered the following explanation of this phenomenon:

> Now that it has been empirically shown that convictions are better predictors of accidents than accidents are of themselves, inquiry into the possible reason for the greater stability of convictions leads to a consideration of characteristics which influence stability. Probably the most important factors contributing to the superiority of convictions as a predictor of total reported accidents are their greater frequency of occurrence as compared to accidents and the inclusion of accident related elements in citation frequency. It is this combination of increased stability and the intrinsic overlapping of behavioral elements (between accident behavior and violation behavior) which makes citation frequency a better predictor of accidents.

The above observation and the results of Harano (1975) provided much of the impetus for the present study. The present effort was designed to further explore the viability of predicting total accidents from equations developed to predict total convictions for the general driving population. In addition to being of substantive and theoretical interest, equations or models that identify drivers at increased risk of future accident involvement have practical applications for driver licensing agencies responsible for identifying and controlling negligent or high-risk drivers (Gebers & Peck, 1987).

METHODS

Subjects

Data for the analyses were obtained from the driving records of a 1% random sample of licensed California drivers ($n = 246{,}600$) extracted in 1992 from the department's driver license (DL) master file. Detailed information on this database is provided by Peck, McBride, and Coppin (1971), Peck and Kuan (1983), Peck and Gebers (1992), Gebers and Peck (1987 & 1994), and Gebers (1998 & 1999).

To be eligible for inclusion in the sample, individuals had to possess a valid California driver license as of the beginning of the study period. All drivers with a deceased indicator on their record or whose driver license had been expired for more than 6 months as of the 1992 data extract date were deleted from the sample.

Predictor Variables

The predictor variables are listed below. Variables listed under "A" are licensing and driver-record variables specific to individual subjects. These represent the majority of potentially relevant driving population parameters contained in California driver record files. They were chosen to be consistent with variables used in previous California driver record studies. Variables listed under "B" are territorial variables. These represent variables aggregated by ZIP-Code of driver residence. The first six variables (i.e., % Black through median annual household income) originated from the 1990 U.S. Census. The last two variables (i.e., mean ZIP-Code total citations and mean ZIP-Code total accidents) originated from the department's DL database. These

variables simply represent the citation and accident rates of California drivers residing in each ZIP-Code area. The territorial variables were chosen according to the criteria specified in a technical memo by DeYoung (1993).

A. Licensing and driver-record variables

- Gender (0 = man; 1 = woman)
- Age (at the beginning of the criterion period)
- Prior 3-year total accidents as defined below
- Prior 3-year total citations as defined below
- Possession of a commercial driver license (0 = no; 1 = yes)
- Presence of a physical or mental (P&M) condition (e.g., lapses of consciousness, mental condition, drugs) on record (0 = no; 1 = yes)
- Presence of a driver license restriction on record (0 = no; 1 = yes)

B. Territorial variables within ZIP-Code of residence

- % Black
- % Hispanic
- % on public assistance
- % unemployed
- % age 55 or older
- Median annual household income ($)
- 3-year (1989-91) mean ZIP-Code total citations
- 3-year (1989-91) mean ZIP-Code total accidents

Criterion Variables

Total accidents and total citations during a 3-year subsequent period were the criterion measures used in the analyses.

The accident data represent reported accidents only. California Vehicle Code Section 16000 requires the driver of every motor vehicle involved in an accident resulting in damage to the property of either party 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 suspension of the driving privilege. Accidents involving an injury or fatality must also be reported to the DMV by the California Highway Patrol.

It should be noted that the term "accidents" is used here to mean "accident involvements." More than one driver can be (and indeed usually is) involved in any given accident. If a driver in the 1% random sample collided with another driver from within the same sample, this would be counted as two involvements (one for each driver). Conversely, if a driver in the sample collided with a driver outside of the sample, the accident would count as one involvement.

The total citation count includes citations, failures to appear in court (FTAs), and traffic violator school (TVS) citation dismissals in the defined time period (based on violation date). A citation that was dismissed conditional upon the offender's completion of TVS is counted here even though it is not legally considered a conviction. Each citation incident is counted here as only one conviction, one FTA, or one TVS dismissal, even if

there were multiple violations (e.g., when a driver is cited for speeding and failing to stop for a red light on the same "ticket").

The above relationship between prior and subsequent driving record is referred to here as a nonconcurrent relationship. A nonconcurrent relationship is one in which a criterion variable (e.g., subsequent total accidents) is correlated with a variable measured during a prior period of time (e.g., prior citations). In the following sections, a series of 3 x 3 nonconcurrent analyses will be presented in which the subsequent 3-year driver record is predicted from the immediately preceding 3-year driver record.

Data Analysis

Both multiple linear regression analysis and canonical correlation analysis techniques were used to identify the combination of variables that provided the most accurate prediction of the criterion measure. SAS statistical software PROC CANCORR and PROC REG were used for the statistical analyses (SAS Institute Statistical Software Inc. Version 6, 1990 a and b). Variables significant at .10 alpha ($p < .10$) were candidates for inclusion in the equation.

In multiple regression analysis, the predictor variables are on one side of the equation, and a single dependent variable (e.g., accident involvements) is on the other side. The predictor variables are weighted and combined to yield a predicted value that maximizes the correlation between the predicted value and the single dependent variable.

Canonical correlation analysis is similar to multiple regression analysis except that there are several variables on both sides of the equation. A canonical correlation analysis is a multivariate regression technique in which a set of two or more dependent variables is regressed against a set of independent or predictor variables. Variables on each side of the equation are optimally weighted and combined in a linear fashion to produce the highest correlation between the two variable sets.

The rationale for using both techniques in the study was to determine if the identification of future accident-involved drivers can be further improved by use of canonical analysis.

For cross-validation purposes, two samples (a construct sample and a cross-validation sample) were generated based on driver license number for both the multiple regression and the canonical correlation analyses. Drivers with an odd sixth-digit of the license number were assigned to the construct sample, and drivers with an even sixth-digit of the license number were assigned to the cross-validation sample. Regression weights (coefficients) derived from the construct-sample equations were applied to the cross-validation sample to test their validity. Validity coefficients were computed by correlating the actual and predicted criterion values for the cross-validation sample.

It should be noted that the multiple regression and canonical correlation techniques use the standard ordinary least squares method of estimation, which makes certain assumptions about the data being analyzed. These assumptions do not hold for accident involvement counts, which have a Poisson-like distribution. However, analyses of the same data set by logistic regression and Poisson regression methods

reported in Gebers (1998) indicate that these procedures nevertheless yield almost identical results. In addition, the classification tables produced by the equations were evaluated here by non-parametric techniques (i.e., methods of hypothesis testing such as the phi coefficient and chi square), which are valid under less restrictive assumptions than are parametric ordinary least square techniques.

RESULTS

Sample Characteristics

The similarity between the construct and cross-validation samples was verified by comparing the two groups on several biographical and prior driver record variables. Distributions of means and proportions were obtained for the individual licensing and driver record variables and the territorial variables. Table 1 displays the biographical and prior 3-year driver record variables for both samples.

Table 1

Descriptive Measures for the 6-Year Construct and Cross-Validation Samples

Group attribute	Construct sample (n = 76,194)	Cross-validation sample (n = 76,737)
Licensing and driver record variables		
% male	52.44	52.30
Mean age	45.70	45.64
Mean prior 3-year total accidents	0.170	0.171
Mean prior 3-year total citations	0.647	0.636
% with commercial license	3.33	3.33
Mean driver license restrictions	0.353	0.349
% with one or more P&M conditions	1.42	1.36
Territorial variables		
% Black	6.11	6.23
% Hispanic	22.14	22.07
% on public assistance	3.99	3.99
% unemployed	4.10	4.11
% age 55 and above	18.78	18.78
Median annual household income ($)	43,903	43,893
Mean ZIP-Code total citations	0.197	0.197
Mean ZIP-Code total accidents	0.048	0.048

Note. The samples were not significantly different on any of the descriptive measures ($p > .05$). A t-test was used in the case of continuous variables such as age. A χ^2 test was used in the case of percentages, expressed as frequencies.

No statistically significant differences or biases were found. This makes it very improbable that the cross-validation findings presented below are attributable to initial differences between the two groups.

Predicting Total Accidents

Table 2 displays the results of the multiple regression analysis on subsequent total accidents. A multiple *R* of .158 was generated for the construct sample.

Table 2

Summary of Standard Multiple Regression Analysis for Predicting 3-Year (1989-91) Total Accidents (Construct Sample: *N* = 76,194)

Criterion variable (1989-91)	Predictor variable (1986-88)	Regression coefficient	Standard error	*t*	*p*
Total accidents	Constant	0.133	0.009	15.30	.000
$\overline{X}$ = 0.1515	Prior 3-year citations	0.027	0.001	20.66	.000
SD = 0.4140	Prior 3-year accident involvements	0.054	0.003	15.61	.000
	Possession of commercial license	0.123	0.008	14.71	.000
	Age	-0.001	0.000	-11.08	.000
	Gender	-0.029	0.003	-9.47	.000
	% Black	0.121	0.014	8.39	.000
	% Hispanic	0.071	0.010	7.37	.000
	Median annual household income	56E-8	12E-8	4.67	.000
	Presence of P&M condition on record	0.050	0.013	4.01	.000
	Presence of driver license restriction on record	0.008	0.003	2.30	.022
F for the equation = 197.64					
Cross-validation *r* = .161					
R^2 = .025					
p = .000					

The signs (positive or negative) of the regression coefficients indicate that increased accident involvement is associated with:

- Increased prior citation frequency
- Increased prior accident frequency
- Having a commercial driver license (which is mostly held by high-mileage professional drivers)
- Being young
- Being male
- A higher percentage of Blacks residing within a ZIP-Code area
- A higher percentage of Hispanics residing within a ZIP-Code area
- A higher median income within a ZIP-Code area
- Having one or more P&M conditions on record
- Having one or more driver license restrictions on record

Application of the construct equation to the cross-validation sample resulted in a significant cross-validation correlation coefficient of .161 ($p < .001$). The lack of shrinkage in this measure in the present analysis is due to the large sample and to the low ratio of independent variables to the sample size. According to Pedhazur (1982), one would expect shrinkage from the construct sample regression coefficient to the cross-validation sample correlation coefficient of no more than .01 when this ratio is 1:50 or less. The 1:50 ratio is much greater than the ratio of about 1:7,619 for these data.

The ability of the equation to predict subsequent accident involvements in the cross-validation sample is illustrated in Table 3. In this table, predicted accident counts are cross-tabulated against the actual (observed) accident counts recorded for subjects in the cross-validation sample. The "cut-point" for using the equation to classify drivers into 0 vs. 1 or more accident groupings has been selected to roughly equalize the marginal frequencies. Equalizing the marginal frequencies produces an equal number of false-negative and false-positive errors (i.e., the unshaded cells in Table 3). The accuracy of the equation in predicting accident involvement can be determined by comparing the predicted scores to the actual accident involvement. The results indicate that the equation significantly discriminated between accident-involved and accident-free drivers ($\chi^2 = 917.48$, $p < .001$). However, the ability of the equation to correctly predict an accident-involvement outcome is low, as evidenced by the 22.8% (2,331 ÷ 10,228) accuracy rate (true positives). The total percentage correct (79.4%) and the accuracy of the accident-free predictions appear to be high, but this is largely attributable to the fact that the great majority of drivers (86.7%) were accident-free during the 3-year period.

Table 3

Actual Total Accidents by Predicted Total Accidents
(Cross-Validation Sample: $N = 76{,}737$)

	Predicted total accidents					
Actual total accidents	0	1 or more	Row total	Percent of N	Correct classifications as percentage of row total	Correct classifications as percentage of grand total
0	58,621	7,888	66,509	86.67	88.14	
1 or more	7,897	2,331	10,228	13.33	22.79	
Column total	66,518	10,219	76,737	100.00		79.4

Note. $\chi^2 = 917.48$ ($p < .001$); phi coefficient = .109. Shaded boxes represent correct classifications. Cut-off scores were established to approximate marginal totals.

The phi coefficient given in the table footnote can be interpreted as a Pearson r for binary data and is sometimes referred to as the point biserial correlation. The absolute

value of the phi coefficient can vary between 0 and 1; the larger the value, the stronger is the relationship between the two variables. As would be expected from the high proportion of false-positives, the phi coefficient is low (.109), indicating that the equation has only a very modest ability to predict accident involvement.

Predicting Total Citations

Table 4 summarizes the results of the citation-prediction equation. In predicting subsequent citations, a construct R of .453 and a cross-validation r of .454 were calculated.

Table 4

Summary of Standard Multiple Regression Analysis for Predicting 3-Year (1989-91) Total Citations (Construct Sample: N = 76,194)

Criterion variable (1989-91)	Predictor variable (1986-88)	Regression coefficient	Standard error	t	p
Total citations	Constant	0.830	0.020	41.64	.000
$\overline{X}$ = 0.5584	Prior 3-year citations	0.287	0.003	94.53	.000
SD = 1.0496	Age	-0.011	0.000	-46.45	.000
	Gender	-0.215	0.007	-30.59	.000
	Prior 3-year accident involvements	0.088	0.008	11.12	.000
	Possession of commercial license	0.156	0.019	8.12	.000
	% Black	0.197	0.033	5.98	.000
	Presence of P&M condition on record	0.122	0.029	4.25	.000
	Median annual household income	157E-8	28E-8	5.61	.000
	% Hispanic	0.112	0.022	5.08	.000
	Presence of driver license restriction on record	-0.016	0.008	-2.14	.032
F for the equation = 1969.66					
Cross-validation r = .454					
R^2 = .205					
p = .000					

The signs of the regression coefficients indicate that an increased number of citations is associated with:

- Increased prior citation frequency
- Being young
- Being male
- Increased prior accident frequency
- Having a commercial driver license
- A higher percentage of Blacks residing within a ZIP-Code area
- Having one or more P&M conditions on record
- A higher median income within a ZIP-Code area

- A higher percentage of Hispanics residing within a ZIP-Code area
- Absence of a license restriction code on record (note reversal in direction from the coefficient in the accident-prediction model)

The accuracy of the construct regression equation in predicting subsequent citations is illustrated in Table 5. The significant χ^2 value of 15,584 ($p < .001$) indicates that the equation significantly discriminated between citation-involved and citation-free drivers.

Because the phi-coefficient is not applicable to a contingency table with more than four cells, a different measure (contingency coefficient) was used to calculate the magnitude of the association between actual and predicted citation frequencies in Table 5. This measure, symbolized by *C*, produced an index of .411, which is relatively close to the Pearson validity coefficient ($r = .454$) for the multiple regression equation. Although these coefficients are of moderate size, they are much larger than those produced for the accident equation. This result is consistent with the prior literature and with the fact that accidents are less predictive and more affected by external stochastic and random influences than are citations.

Table 5

Actual Total Citations by Predicted Total Citations
(Cross-Validation Sample: $N = 76{,}737$)

	Predicted citations								
Actual citations	0	1	2	3	4 or more	Row total	Percent of *N*	Correct classifications as percentage of row total	Correct classifications as percentage of grand total
0	39,550	8,628	2,167	613	339	51,297	66.85	77.10	
1	8,886	4,432	1,631	612	378	15,939	20.77	27.81	
2	2,080	1,783	913	405	362	5,543	7.22	16.47	
3	556	677	431	217	255	2,136	2.78	10.16	
4 or more	226	441	384	301	470	1,822	2.37	25.80	
Column total	51,298	15,961	5,526	2,148	1,804	76,737	100.00		59.4

Note. $\chi^2 = 15{,}584$ ($p < .001$); $C = .411$. Shaded boxes represent correct classifications. Cut-off scores were established to produce approximately equal row and column marginal frequencies.

A comparison of the two classification matrices can best be done by collapsing Table 5 into a 2 x 2 table in which the predictions and actual values are in terms of 0 versus 1 or more. The phi coefficient and percentage of drivers getting a citation who are accurately classified are, respectively, .309 and 53.82%. These indices are substantially higher than the respective figures for accidents shown in Table 3.

Predicting Accident Involvement Using The Citation Equation

The results presented above indicate that the accident equation was only marginally successful in predicting accident involvement. Consequently, an attempt to improve

prediction was made by using predicted citations rather than the accident equation itself to identify drivers involved in subsequent accidents. It was hypothesized that accidents may be predictable through the citation equation, as accidents and citations are known to have shared causative factors.

In an attempt to evaluate this hypothesis, actual accident involvement of the sample was cross-tabulated against predicted citation involvement. In other words, we are illustrating the extent to which drivers who are predicted to be citation-involved will also be accident-involved during that same time period. The results are presented in Table 6. In this table, drivers predicted to have two or more subsequent citations were predicted to be accident-involved, while drivers predicted to have fewer than two citations were predicted to be accident-free. These particular cut-off points were established to approximate equal row and column marginal frequencies. The table displays the actual accident status of the individuals predicted to be accident-involved under this schema.

The statistically significant χ^2 value of 800.92 ($p < .001$) indicates that subsequent total accidents could be significantly predicted by the citation equation. Note also that the phi coefficients in Tables 2 and 6 are almost identical (.109 vs. .102), indicating that in practical terms the two equations perform similarly in identifying accident-involved drivers. However, the small difference in the phi coefficients is statistically significant ($p < .01$), indicating some reduction in the classification accuracy of the conviction-mediated equation compared to the accident equation. This can be seen by comparing the respective percentage of accident-drivers who were correctly classified by the two equations (20.9% for the citation mediated equation vs. 22.8% for the accident mediated equation).

Table 6

Actual Total Accidents by Predicted Total Citations
(Cross-Validation Sample: $N = 76{,}737$)

	Predicted citations					
Actual total accidents	0 or 1	2 or more	Row total	Percent of N	Correct classifications as percentage of row total	Correct classifications as percentage of grand total
0	59,171	7,338	66,509	86.67	88.97	
1 or more	8,088	2,140	10,228	13.33	20.92	
Column total	67,259	9,478	76,737	100.00		79.90

Note. $\chi^2 = 800.92$ ($p < .001$); phi coefficient = .102. Shaded boxes represent correct classifications. Cut-off scores were established to approximate equal marginal totals. The correlation coefficient between the number of actual total accidents and the number of predicted citations is .144.

Predicting Total Accidents Using A Multivariate Equation

As noted earlier, the above analyses use multiple regression as the analytical tool. The multiple regression equation was used to explain, or predict, either total accidents or total citations on the basis of multiple independent variables. In this section, the results from a series of canonical correlation analyses show the relationships between the

multiple independent variables and the dependent variables (accident and conviction involvements) in combination.

Specifically, canonical correlation analysis was used to predict a vector of subsequent total accidents and citations from the set of independent variables. The two canonical functions or roots obtained in each analysis were used to classify drivers in a series of 2 x 2 tables. This enabled comparisons to be made with the tables produced from the separate accident and citation regression equations presented above.

Table 7 summarizes the canonical correlation results for the nonconcurrent 6-year construct sample.

The canonical correlations (R_{root1} and R_{root2}) are displayed in the bottom of Table 7. The first canonical correlation is .4577, indicating 20.95% (i.e., $.4577^2$) overlapping variance for the first pair of canonical variates. The second canonical correlation is .0762, indicating 0.58% (i.e., $.0762^2$) overlapping variance for the second pair of canonical variates. Although highly significant overall ($F = 975.10$, $p < 0.001$), neither of these two canonical correlations represents a strong relationship between pairs of canonical variates.

Table 7

Summary of Nonconcurrent 6-Year (1986-87; 1989-91) Canonical Correlation Results (Construct Sample: $N = 76{,}194$)

Independent variables	Root 1		Root 2		Dependent variables	Root 1		Root 2	
	B	*S*	*B*	*S*		*B*	*S*	*B*	*S*
Prior total citations	0.722	.898	-0.306	-.150	Subsequent total accidents	0.143	.309	1.005	.951
Prior total accidents	0.095	.298	0.602	.583	Subsequent total citations	0.965	.990	-0.312	-.141
Commercial license class	0.073	.181	0.595	.619					
Age	-0.357	-.581	0.103	.124					
Gender	-0.227	-.396	-0.041	-.111					
% Black	0.052	.071	0.330	.312					
% Hispanic	0.050	.088	0.329	.270					
Median income	0.053	-.010	0.173	-.114					
P&M code	0.034	.076	0.133	.143					
Restriction status	-0.013	-.207	0.144	.180					
PV:		.148		.101			.538		.462
Rd:							.113		.003
Total Rd:								.116	
$R_{root\ 1} = 0.4577$ $R_{root\ 2} = 0.0762$									

Note. *B* = standardized coefficient; *S* = structure or loading coefficient; PV = proportion of variance extracted; Rd = redundancy; Total Rd = total redundancy. The *F* value for both canonical variate pairs is 975.10 ($p < .001$). The *F* value of the second canonical variate after "peeling off" the first canonical variate pair is 49.46 ($p < .001$).

As stated above, both canonical correlations are statistically significant and therefore are considered to be different from zero. This result is to be expected from the fact that both dependent variables are known to be related to many of the independent variables used in this study—as evidenced by the preceding multiple regression results. The fact that the second function accounts for such a small percentage of variance (0.58%) after the first function has been extracted indicates that the first function is by far the more important of the two.

Although the canonical correlations provide a measure of shared or overlapping variance, they do not represent the percentage of variance in the dependent variable vector (accidents and convictions) that can be predicted or explained by the vector of independent variables. In canonical correlation analysis, this latter index, which is analogous to R^2 in multiple regression, is provided by the redundancy statistic. Note from Table 7 that this index for the largest function and for both functions combined is, respectively .113 and .116. Thus, the two functions explain only 11.6% of the variance on the dependent variable vector.

The structure or loading coefficients are presented in the column labeled *S* in Table 7. A structure or loading coefficient is the correlation between a given original variable (not combined with others) and the canonical variate scores. As a rule of thumb, some authorities recommend that only coefficients with an absolute magnitude of .30 or higher be treated as meaningful (Pedhazur, 1982, Tabachnick & Fidell, 1996). Using this guide for the criterion vector of total citations and accidents, one would conclude that both total citations and total accidents have meaningful loadings on the first canonical variate but that only total accidents has a meaningful loading on the second canonical variate. Since the goal of this analysis is to develop the most efficient multivariate model for the prediction of subsequent accidents, the remaining discussion will focus on the first canonical variate.

Using the .30 guideline stated above for interpretation of the structure coefficients, an examination of these coefficients under the two columns labeled Root 1 indicates that the multivariate vector of subsequent traffic incidents (i.e., accidents and citations) is associated with increasing counts of prior total citations, increasing counts of prior total accidents, being young, and being male.

The ability of the canonical variate pair (Root 1) to predict actual subsequent accident involvement in the cross-validation sample is presented in Table 8. In the table, the predicted canonical variate scores are cross-tabulated against the actual, observed accident counts for the cross-validation sample. Scores on the canonical variates were calculated as the product of drivers' standardized scores on the original variates of the total accidents and total citations, weighted by the canonical coefficients from Table 7 (0.143 and 0.965, respectively).

Table 8

Actual Total Accidents by Predicted Driving Incidents—Accidents and Citations (Canonical Variate Scores) (Cross-Validation Sample: $N = 76{,}737$)

	Predicted driving record incidents (canonical variate score)					
Actual total accidents	0	1 or more	Row total	Percent of N	Correct classifications as percentage of row total	Correct classifications as percentage of grand total
0	59,344	7,165	66,509	86.67	89.23	
1 or more	7,552	2,676	10,228	13.33	26.16	
Column total	66,896	9,841	76,737	100.00		80.82

Note. $\chi^2 = 1{,}878.20$ ($p < .0001$); phi coefficient = .156. Shaded boxes represent correct classifications. Cut-off scores were established to approximate equal marginal totals.

The results indicate that the canonical variate pair (a function of independent variables predicting a function of dependent variables) significantly discriminated between accident-involved and accident-free drivers ($\chi^2 = 1{,}878.20$, $p < .0001$). The canonical correlation technique was substantially superior to the other accident prediction strategies (see Tables 3 and 6) as evidenced by the phi coefficient of .156 compared to the previous phi coefficients of .109 and .102. The ability of the canonical equation to correctly predict accident involvement was also higher than that of either the citation or accident equation alone, as evidenced by a 26.16% accuracy rate in identifying accident-involved drivers. This compares to the 22.79% accuracy rate produced by the accident equation, representing a 14.8% increase in predictive accuracy.

DISCUSSION

Although the results do not support the hypothesis that equations keyed to citations do as well as or better than equations keyed to accidents in predicting subsequent accidents, the fact that equations keyed to citations identify groups who are almost as likely to be accident-involved as are drivers identified by the accident equation is noteworthy. The negligent-driver point systems of most states are weighted more by citations than by accidents. As a result, drivers are much more likely to receive DMV license controls due to citations than to accidents. The fact that drivers being treated by license control programs based on the point system are also highly involved in accidents suggests that a program that targets conviction repeaters may be close to optimal in terms of targeting accident-prone drivers.

The results suggest that identification of future accident-involved drivers can be improved by either of two approaches. The first is to construct equations based on a combination (perhaps a simple sum) of prior accidents and citations. To some extent, California's neg-op point system reflects such an approach since points are allocated to

traffic convictions and culpable accidents. The second alternative is more elaborate, involving a truly multivariate approach in which the prediction equation consists of a two-variable vector of subsequent citations and accidents. The canonical correlation analysis performed for this study resulted in two orthogonal canonical functions or roots: A driving-incident function consisting of primarily citations and secondarily accidents and an almost exclusively accident function. The second function was disregarded because it accounted for only a negligible increase in explained variance. The canonical variate score produced from the first function resulted in an accident "hit rate" that was significantly higher than those of the preceding equations.

The extraction of two canonical variates in this study is consistent with intuition and substantive theoretical considerations. Traffic conviction frequency is known to be correlated with increased accident propensity and reflects both risk-taking, social nonconformity, and exposure. However, accidents can also be associated with other individual differences among drivers, such as driving skill, information processing ability, and level of cognitive functioning. While it is true that accidents and citations share a number of common overlapping elements, the fact that the two canonical functions' respective variables contain loading coefficients of different signs and magnitudes provides evidence for the hypothesis that citations and accidents contain some idiosyncratic variance. The first function appears to capture variance in accident propensity related to citations and citation-accident covariation whereas the second function reflects variance in accident propensity which is unrelated to citations.

Inspection of the structure loadings of the two functions yields some additional insights. The highest loading for function 2 was on the commercial license class variable. It therefore appears that the second function is mediated by license class which, when considered along with the loading on prior accidents, suggests a function that distinguishes accident-involved commercial drivers from accident-free commercial drivers.

As noted above, the results presented in this paper contradict the findings of Harano (1975) and Marsh and Hubert (1974). These authors found, in separate studies, that a multiple regression equation generated to predict subsequent citations could predict subsequent accidents as well as or better than an accident-prediction equation when applied to a cross-validation sample. The failure to replicate the earlier findings is probably due to the differences in the study populations. The sample used in the present effort consisted of a random sample of all California drivers. However, the samples used by Harano (1975) and Marsh and Hubert (1974) consisted only of negligent drivers. In fact, the sample utilized by Marsh and Hubert was restricted to a sample of male negligent drivers. Negligent drivers, in addition to representing a much more homogenous group, would differ dramatically from general driving populations on a number of characteristics.

Although the findings show that the accuracy of the prediction models greatly exceed chance expectations, the best model had only a 27.2% accuracy in predicting which drivers would be accident-involved during the subsequent 3-year period. Thus, 72.8% of the drivers predicted to be accident-involved remained accident-free. Stated another way, 72.8% of the subsequent accidents involved drivers who were predicted to be free of accidents. It would be possible to increase the specificity of the accident predictions

by altering the cut-off value used to classify drivers into the accident-free vs. accident-involved predicted dichotomy. For example, rather than predicting 13% of the sample to be accident-involved, we could use a much higher cut-off threshold, say one which would predict only the "worst" 5% or 1% of the sample to be accident-involved. If this were done, the model would have much greater specificity in that those predicted to be accident-involved would be much more accurately classified (lower false positive rate). However, there would be a reciprocal decrease in the sensitivity of the model—that is, the total proportion of accident involvements which was correctly predicted would be greatly diminished (higher false negative rate). Which type of error to minimize is a complex issue but at an abstract level involves a consideration of the costs or relative disutility of the two errors. Is it, for example, more serious to not take action against a driver who will become accident-involved than it is to impose a driver control action on a driver who would have remained accident-free in absence of the action? To some extent the answer would turn on the nature of the driver control action taken. If the actions were relatively non obtrusive, such as warning letters or informational material, we would be less concerned with false positive errors. However, expensive countermeasures or obtrusive actions like license revocation might require a low to moderate false positive error rate.

It is instructive to consider how the department (implicitly) weighs these trade-offs by considering the "deviancy" thresholds at which license control actions are currently taken in California. Drivers defined as "negligent" in accord with the prima facie definition of the California Vehicle Code represent .90% of the driving population and have a subsequent one year accident rate which is roughly 3.5 times that of point-free drivers. These drivers are subjected to driver control actions, including license suspension. Thus, the department frequently suspends traffic conviction and accident repeaters whose point count exceeds 99% of all drivers. If this deviancy criterion were applied to the model developed herein, a cut-point would be selected for predicting accident involvement that would be exceeded by only 1% of all drivers. This strategy would result in false positive and false negative rates that were dramatically different than those shown in this paper. More specifically, the use of a 99th percentile cut-point would achieve a respectable degree of specificity (false positive rate) but at the cost of greatly reduced sensitivity (false negative).

In conclusion, the results reported in this study, like those of earlier studies, indicate that subsequent driving record can be predicted from prior driving record for groups of individuals but that the error rates at the individual level are inherently large. The model derived from the canonical analysis, while superior to the simpler models, would be very difficult to implement operationally. The most obvious problem relates to its complexity. Canonical correlation analysis is not easy to comprehend, and the task of explaining the meaning of a canonical based point system to administrators and legislators, let alone the public, is daunting if not prohibitive. Another problem is that the equations contain a number of variables (age, gender, etc.) that would not be legally defensible in taking license control actions. This problem could be rectified, with some sacrifice in predictive power, by deleting the unacceptable variables. In addition, use of variables such as age and gender might be permissible for triggering educational or advisory interventions.

REFERENCES

Burg, A. (1967). *Vision test scores and driving record: General findings*. Report No. 67-24. Los Angeles: University of California Los Angeles, Institute of Transportation and Traffic Engineering.

Burg, A. (1975). *Traffic violations in relation to driver characteristics and accident frequency*. Report No. 74-55. Los Angeles: University of California Los Angeles, Institute of Transportation and Traffic Engineering.

Cobb, P. W. (1940). The limit of usefulness of accident rate as a measure of accident-proneness. *Journal of Applied Psychology, 24*, 154-159.

DeYoung, D. J. (1993). *Reduced set of 1990 census variables*. (Internal memorandum to research analysts and managers, December 3). Sacramento: California Department of Motor Vehicles.

Gebers, M. A. (1999). *Strategies for estimating driver accident risk in relation to California's negligent-operator point system* (Report No. 183). Sacramento: California Department of Motor Vehicles.

Gebers, M. A. (1998). *Exploratory multivariable analyses of California driver record accident rates*. *Transportation Research Record, 1635*, 72-80.

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., & Peck, R. C. (1987). *Basic California traffic conviction and accident record facts* (Report No. 114). Sacramento: California Department of Motor Vehicles.

Harano, R. M. (1975). The psychometric prediction of negligent driver recidivism. *Journal of Safety Research, 7*(4), 170-179.

Marsh, W. C., & Hubert, D. M. (1974). *The prediction of driving record following driver improvement contacts* (Report No. 50). Sacramento: California Department of Motor Vehicles.

Newbold, E. (1927). Practical applications of the statistics of repeated events, particularly to industrial accidents. *Journal of the Royal Statistical Society, 90*, 487-547.

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., & Kuan, J. (1983). A statistical model of individual accident risk prediction using driver record, territory and other biographical factors. A*ccident 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.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd Ed.). New York: Holt, Rinehart, & Winston.

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

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

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd Ed.). New York: Harper Collins.