DOT HS-804 851

TECHNIQUES FOR PREDICTING HIGH-RISK DRIVERS FOR ALCOHOL COUNTERMEASURES
Volume I: Technical Report

John H. Lacey
J. Richard Stewart
Forrest M. Council

University of North Carolina
Highway Safety Research Center
Chapel Hill, North Carolina 27514

Contract No. DOT HS-5-01250
Contract Amt. $147,773

MAY 1979
FINAL REPORT

Prepared For
U.S. DEPARTMENT OF TRANSPORTATION
National Highway Traffic Safety Administration
Washington, D.C. 20590

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT HS 804 851		

4. Title and Subtitle	5. Report Date
Techniques for Predicting High-Risk Drivers for Alcohol Countermeasures. Volume I. Technical Report	May 1979
	6. Performing Organization Code
	8. Performing Organization Report No.
7. Author(s)	
John H. Lacey, J. Richard Stewart, Forrest M. Council	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
University of North Carolina Highway Safety Research Center South Campus, CTP 197A Chapel Hill, NC 27514	11. Contract or Grant No. DOT-HS-5-01250
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address U. S. Department of Transportation National Highway Traffic Safety Administration Washington, D.C. 20590	Final Report June 1975 -- June 1978
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This technical report, a companion to the Volume II User Manual by the same name describes the development and testing of predictive models for identifying individual with a high risk of alcohol/related (A/R) crash involvement.

From a literature review and an analysis of North Carolina accident data, six groups of drivers at high risk of A/R crashes were identified: 16 to 20-year-old males; 21 to 24-year-old males; persons with previous DUI convictions; persons with three or more moving violations; persons recently divorced; and persons recently released from prison. Using North Carolina data through 1974, predictive models were developed for each of these groups to predict 1975 A/R accident involvement proportions for subgroups within each high-risk group.

Prospective analyses of the models' predictive capabilities using 1976 crash data indicated that, in general, they effectively identify the driver subgroups that have the highest risk of A/R crash involvement. Because most of the information used in the model development is readily available in most governmental jurisdictions, the models can be adapted for use in areas other than North Carolina. The primary drawback to widespread use of the models is the lack of demonstrably effective countermeasure programs that significantly reduce the rate of A/R crash involvement for the identified driver subgroups.

17. Key Words	18. Distribution Statement
Alcohol Alcohol-Related Crashes Crash Prediction High-Risk Groups	Unlimited availability through the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	162	

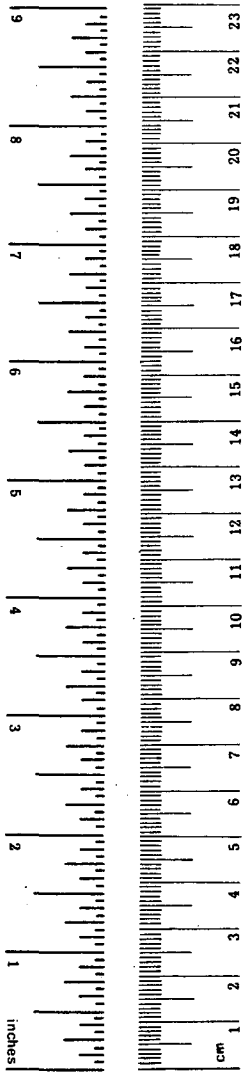
Form DOT F 1700.7 (8-72) Reproduction of form and completed page is authorized

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

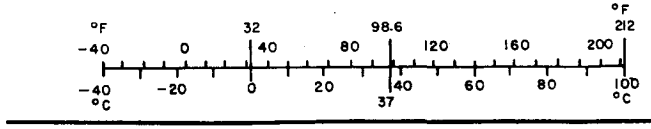
Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
	AREA			
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
	MASS (weight)			
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
	VOLUME			
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
	TEMPERATURE (exact)			
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.



Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
	AREA			
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
	MASS (weight)			
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
	VOLUME			
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
	TEMPERATURE (exact)			
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F



ACKNOWLEDGMENTS

A project with the duration and complexity of this one certainly required the efforts of more than the listed authors. We wish to thank the many persons who made the preparation of this report possible.

Brian Powers and Catherine Mullen assisted greatly in performing the literature review component of the high-risk group selection process.

Virtually every HSRC programmer has worked on this project but we wish particularly to acknowledge the efforts of Fred Diggs, Willie Fischer, Eric Rodgman, and Nancy Woody.

In the preparation of the drafts, briefings and final report, invaluable assistance was offered by Donna Suttles, Peggy James, Teresa Parks and Martha Apple in typing and text preparation, by Cranine Brinkhous, Bill Pope and Lauren Ogle in table, graph, and slide preparation, and by Patricia Waller, Carol Carroll, and David Cole in carefully reviewing the final report. Frank Roediger provided editorial assistance throughout the preparation of the various drafts.

We wish to particularly thank personnel from the Division of Motor Vehicles, Department of Correction and Department of Human Resources for making available to us the data used in developing the models.

Marvin M. Levy was the Contract Technical Manager.

ADDENDUM

In this volume (Volume 1) of the final report, the Contractor found that drivers can be identified who have an extremely elevated risk of being involved in an alcohol-related crash. However, only 8% of those drivers identified as most likely to be involved in an alcohol-related crash are likely to actually become involved in such a crash within the next twelve months. This means that even if a countermeasure was found which was fully effective in preventing alcohol-related crashes, it would be expected to prevent only about 8 crashes for every 100 high risk drivers impacted by that countermeasure. As reported in this document (see Table 7.3, page 114) estimates of countermeasure effectiveness (for reducing alcohol-related crashes) were much more modest; therefore, the screening approach used here for identifying high risk drivers is not likely to result in a large reduction in the occurrence of alcohol-related crashes.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

EXECUTIVE SUMMARY

This report concerns an effort to develop and test a predictive modelling technique to identify individuals at high risk of alcohol/related (A/R) crash involvement prior to crash occurrence. A parallel effort, described in Volume II, User Manual presents ways in which alcohol administrators may use the predictive models developed under this project. The study was done to address a perceived need for developing ways to implement alcohol driving countermeasures so that some of the more serious consequences of alcohol-impaired driving might not occur.

The basic approach followed was to identify several groups of drivers known or suspected to be at a high-risk of A/R crash involvement and then, for each group, to separately develop a statistical model which identifies those individuals within each high-risk group that are at an even higher risk of A/R crash involvement.

Six high-risk groups were identified for study through a literature review and rudimentary analysis of North Carolina accident data. The high-risk groups so identified were males, 16-20; males, 21-24; persons with previous convictions for driving under the influence; persons with three or more moving violations; persons recently divorced; and persons recently released from prison. An examination of N.C. accident data for 1973, 1974 and 1975 revealed that a larger proportion of each of these groups was involved in A/R crashes than the general driving population. A one-tenth sample of the general driving population was also selected for comparative purposes in the model development process. In all, models were developed for seven groups--the six high-risk groups and the sample of the general driving population.

The models were developed using data available through 1974 to predict A/R crashes in 1975. The basic data sources were the N.C. Division of Motor Vehicles Driver History File and Accident Files, a listing of persons divorced in 1974 obtained from the N.C. Department of Human Resources and a listing of persons released from prison in 1972 obtained from the N.C. Department of Correction. The data sources were purposely selected to be ones which would be readily and inexpensively available to program administrators so that the models developed from them could be practically replicated and used in other governmental jurisdictions.

In the model development phase, the data sets above were merged and then were examined for each group to identify those variables most highly related to subsequent A/R crashes for that group. Then for each group a predictive model was developed using a categorical data analysis technique called GENCAT. Using this technique, subgroups within each high-risk group (and the general population sample) were identified and assigned a predicted probability of being involved in an A/R crash in 1975.

For each high-risk group, the subgroups with the highest predicted A/R crash experience, that predicted value and the range of predicted values for the whole group are tabulated below.

Group	Subgroup (Individuals with all the characteristics listed)	Predicted Proportion of A/R Crash Involvement	Range of Predicted Values
General population sample	Some days under suspension or revocation (S/R) Some accident violations Male Some reckless violations	.03600	.00050-.03600
Males, 16-20	Some days S/R Some violations Some night crashes Some night alcohol violations	.05679	.00933-.05679

Group	Subgroup (Individuals with all the characteristics listed)	Predicted Proportion of A/R Crash Involvement	Range of Predicted Values
Males, 21-24	Some days S/R Some reckless violations Some alcohol violations Some previous A/R crashes	.06777	.00698-.06777
DUI	Young Some speeding violations Some days S/R Some reckless violations	.07701	.01507-.07701
3 or more violations	Young Male Some days S/R Some previous A/R crashes Some previous crashes	.06780	.00589-.06780
Divorce group	Some alcohol violations Some reckless violations	.05119	.00570-.05119
Prison group	Some administrative violations Young	.0734	.0184-.0734

The highest risk subgroup was in the DUI group and had a predicted A/R crash experience of .07701. This represents a risk 21 times greater than that of the general driving population as a whole (.00362).

Three different data sets were used in assessing the accuracy of the model predictions. These were: (1) the actual 1975 A/R crash experience of one-third of each group which was not used in the model development phase but reserved for this purpose, (2) the whole groups' 1976 crashes, and (3) the 1976 A/R crash experience of newly identified persons who constituted new high-risk groups identified as of the end of 1975. The analysis of the predictive validity of the models indicated that they were quite effective in identifying which of the subgroups were likely to be at the highest risk of A/R crash involvement.

The potential usefulness of the models in a real world applications setting is also discussed. It is concluded that, because most of the predictor

variables have a strong statistical intuitive relationship with A/R crashes, the models may usefully be applied as a means of identifying persons for entry into countermeasure programs. A difficulty is that an attempt to identify countermeasure programs for use in the users manual (Volume II) revealed few scientifically valid studies which indicated that particular countermeasures were effective. This led to a recommendation that the models might best be currently used in conjunction with well conceived evaluations of A/R crash reduction countermeasures.

CHAPTER 1 - INTRODUCTION

Alcohol has long been recognized as a major factor in highway traffic accidents. As early as 1938 Holcomb reported the presence of alcohol in a higher proportion of drivers in personal injury accidents than in a sample of the general driving population. This overrepresentation was concluded to be indicative of alcohol as a causative factor in motor vehicle accidents. Subsequent research, including the Grand Rapids study of Borkenstein, et al. (1964), has further documented in greater detail and with a higher degree of precision the detrimental influence of alcohol on driving performance as measured by accident involvement. This increasing body of evidence led to the emergence of a variety of countermeasure approaches most often characterized by public information programs urging persons not to drive after drinking.

The important role of alcohol in highway crashes was further clarified by the 1968 Department of Transportation report to Congress on alcohol and highway safety which summarized the results of many studies on the subject. One of the findings noted was that there was little known concerning the effectiveness of the various countermeasures attempted to date. A recommendation was that further countermeasure research activity be funded in demonstration and evaluation projects.

The most visible of these new attempts to affect alcohol-related crashes was the federally funded Alcohol Safety Action Project (ASAP) program. This ambitious effort involved the implementation of coordinated multi-pronged countermeasure approaches to the alcohol-related crash problem in selected communities throughout the country. In all, 35 ASAP projects were funded by the federal government at a cost of some $88,000,000.

Since each of these projects used a variety of countermeasure approaches, it proved difficult to assess the effectiveness of any one countermeasure. However, NHTSA has argued that, overall, the program was effective (U.S. DOT, 1974), largely on the basis that the ratio of nighttime to daytime crashes in the project areas decreased after implementation. The rationale here is that nighttime crashes are those most likely to be alcohol-related and thus impacted by the programs, and that the daytime crashes serve as a control. This evaluation approach has not satisfied all critics (Zador, 1976) and efforts are continuing to refine the evaluation of the ASAP projects (Monaco, 1977). However, effectiveness aside, one point not at issue is that the ASAP type approach to reducing alcohol related crashes is an expensive one.

Thus, with the limited amount of highway safety funds available and the wide variety of highway safety needs to be addressed, there is a need to identify a more focused approach to combat the alcohol-related (A/R) crash problem. The research discussed in the remainder of this report pursued one such approach -- to examine the feasibility of identifying individuals or groups of individuals at extremely high-risk of A/R crash involvement so that they might be brought into countermeasure programs. The project addressed two basic questions:

1. Can individuals at high-risk of alcohol-related crash involvement be identified before they have an A/R crash; and
2. Can effective countermeasures appropriate to such individuals be identified from currently available information?

To answer the first question the following approach was used. Several groups of drivers were identified (through a review of the literature and a preliminary analysis of North Carolina data) as being at a high-risk of A/R crash involvement. Predictive models using multivariate techniques to identify subgroups at even higher A/R crash risk were developed for each of the high-risk

groups. The validity in predicting A/R crash experience for each of the models developed was then determined for both a control group from the same time period as the data used to develop the models and for a subsequent year's crash experience. The work described above is reported in this Volume I of the report.

In addition to the model development and testing efforts, a second related project goal was the development of a methodology designed to aid alcohol program administrators in (1) using the developed models to choose high-risk drivers for treatment, (2) selecting an appropriate countermeasure for those drivers using information on cost, effectiveness, potential target groups and length of countermeasure effect, (3) determining whether the costs of a given countermeasure will be less than the benefits derived from it; and (4) conducting well-designed evaluations of the countermeasure activities selected in order to establish levels of effectiveness. This methodological process is described in detail in the companion Volume II: User Manual. Its basic components are a set of tables providing the probability of a subsequent A/R crash as predicted by the models, a series of discussions of the content and effectiveness of various potentially useful countermeasure treatments based on a review of current literature, a computerized cost effectiveness methodology to help assess potential treatment payoff for a chosen group of drivers, and an overview of the components which are basic to the evaluation of any A/R countermeasure program.

Subsequent chapters in Volume I present the methodological framework followed in carrying out the project, the selection of the high-risk groups, the data sources and the data processing, the model development, the validity testing, and the conclusions drawn.

CHAPTER 2 - METHODOLOGY

2.1 Introduction

In this chapter the basic steps undertaken in carrying out the project objectives are presented and briefly discussed. Because the intent is to provide an overview of the project framework, detailed discussions of each step are covered in later sections, and not in this chapter. Figure 2.1 is a flow chart of the major task sequence. The two major project goals are 1) high-risk group selection, model development, and testing; and 2) development of a user manual to guide in the implementation of the models in the field. They were parallel and joint efforts and are depicted as such on the flow chart. Subheadings in this chapter are keyed to the boxes in Figure 2.1.

2.2 High-Risk Group Selection

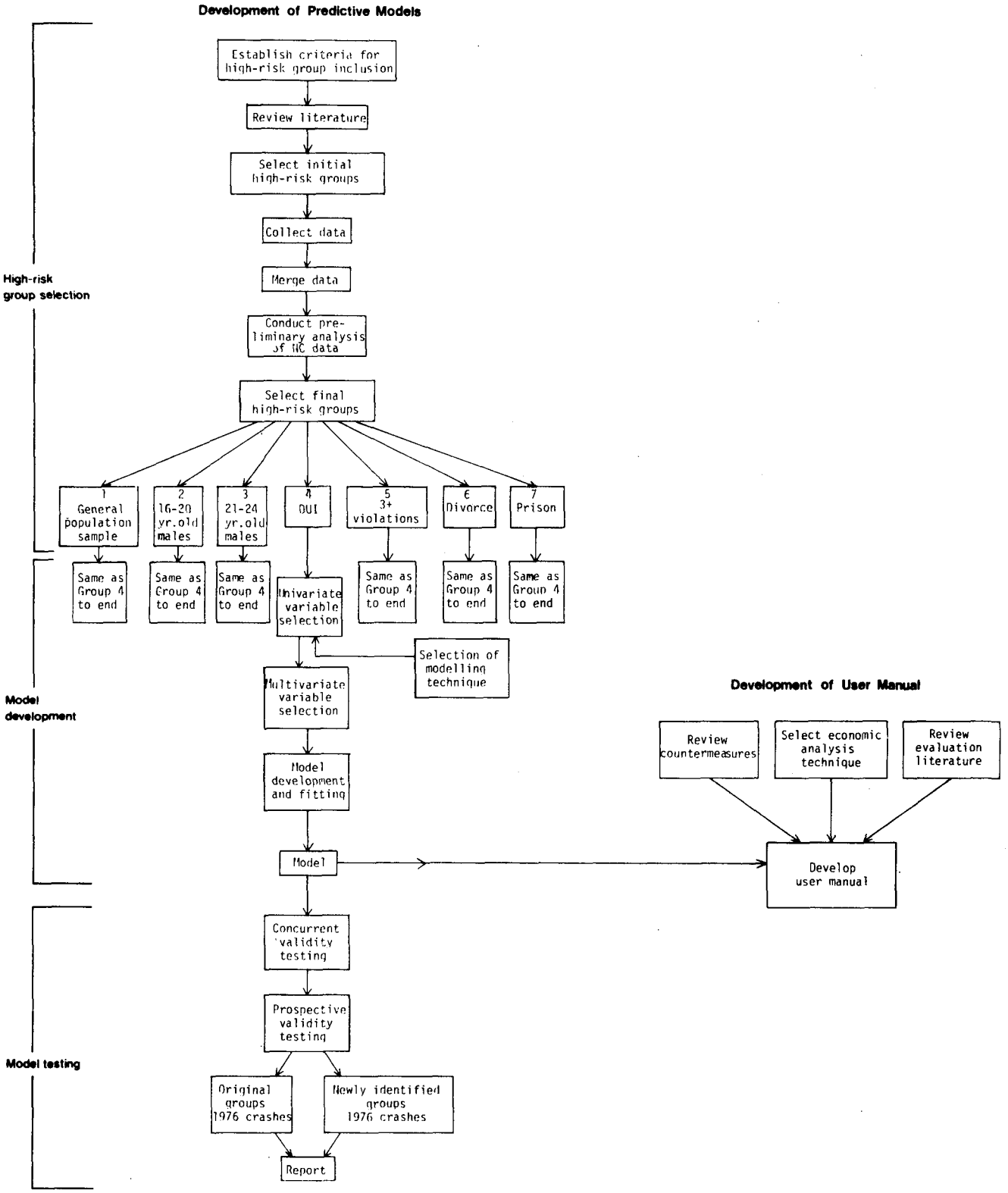
High risk, for the purposes of this study, is defined as an elevated likelihood of involvement in an A/R crash. Thus a high-risk group would be a group of drivers in which a disproportionate share of its drivers subsequently become involved in A/R crashes.

For this study an A/R crash is defined using two variables, "Sobriety" and "Chemical Test Given," which appear on the North Carolina accident report form for every reportable accident. These two variables and the values they may take are shown below:

Table 2.1 Accident report variables used in determination of A/R crashes.

Sobriety	Chemical Test Given
1. Had not been drinking	Yes
2. Drinking - ability impaired	No
3. Drinking - unable to determine impairment	
4. Unknown	

Figure 2.1 Flow Chart of Major Task Sequence



A crash is considered alcohol-related (A/R) if the investigating officer indicated either "Drinking--ability impaired" or "Drinking--unable to determine impairment" in the sobriety variable or indicated that a chemical test was given under the second variable. It was decided to use this liberal definition of an A/R crash to maximize the number of A/R crashes available for the modelling task rather than a more conservative one such as insisting on a determination of impairment. In North Carolina, approximately eight percent of crashes meet the more liberal criterion while only four percent would meet the more conservative one.

2.2.1. Criteria for group selection.

The primary objective of the project was to identify several high-risk groups and develop predictive models which would point out subgroups at even higher risk so that countermeasures might be applied to them. The practical orientation of this project, as evidenced by development of the User's Manual, dictated certain criteria for the high-risk groups.

First, there had to be a reasonable basis to expect the groups to be at an elevated risk of A/R crash involvement. Second, the groups had to be easily and clearly definable. Vague descriptive adjectives such as problem drivers would not suffice. Specific characteristics such as age, sex, recorded driving behavior and the like were considered more appropriate.

Third, the data elements by which the groups were to be defined were to be ones which would be practically available to alcohol program administrators. For the most part, this means that the information should reside in the Motor Vehicles department or in other state agencies. Since the objective of the modelling process is to identify individuals to bring into countermeasure programs, it was assumed that accurate questionnaire type data (such as psychological inventories) could not be obtained from individuals who

might feel that some facet of their driving privilege was at stake. Another consideration in variable selection was the privacy rights of individuals. Though some information such as alcohol treatment center records might be available for research purposes it is doubtful that later, when the models were to be applied for countermeasure purposes, they would remain available to the administrator.

A further consideration in variable selection was that of cost. Even if unbiased questionnaire information were obtainable, the high cost of collecting such information could well render countermeasure programs less cost-effective.

Consequently, the variables selected for high-risk group definition and modelling attempts were restricted to Division of Motor Vehicle records and other computer usable records available on a statewide basis. Thus high-risk groups were to be clearly definable on the basis of information that would be available to Motor Vehicles Administrators at a reasonable cost.

2.2.2. Review literature.

With the criteria outlined above in mind, highway safety and other relevant literature was reviewed in an effort to identify several potential high-risk groups for consideration for inclusion in the modelling process. An effort was made to identify studies which addressed A/R crash risk specifically. However, few studies address that specific issue for subgroups of the driving population while many address crashes, and many others address aberrant drinking behavior of special populations. Thus, in some cases, for high-risk group selection purposes, the logical link between demonstrated aberrant behavior in both driving and drinking was made by the reviewers.

2.2.3. Collect data.

Once preliminary groups were selected based on the review of the literature, the data sources necessary to define the groups were identified and secured.

These included: the Division of Motor Vehicles driver history file, a file of reportable North Carolina traffic accidents, a listing of persons granted a divorce in North Carolina from the N.C. Department of Human Resources, and a file of persons released from prison in N.C. from the N.C. Department of Correction.

2.2.4. Merge data.

The four data sources were of differing types of formats and orientations, but needed to be merged into one file to conduct the study. To that end the divorce and prison files were each ordered alphabetically, computer matched and then hand matched with the driver history file to obtain driver license numbers. The accident files were ordered by driver license number and then all four files were merged into one large file. This file was then broken down for the further analysis into several smaller files corresponding to the high-risk groups.

2.2.5. Conduct preliminary analysis.

For each of the candidate high-risk groups, an analysis of the N.C. data was made in order to determine its appropriateness for further modelling. Each group was analyzed in terms of the percentage of total N.C. A/R crashes it accounted for in 1973, 1974, and 1975. This percentage was termed the impact index. The groups were also examined in terms of their annual A/R crash rate. A risk index was computed which represented the quotient of the annual population A/R crash rate for a given group divided by the general driving population's annual A/R crash rate.

2.2.6. Select final high-risk groups.

Based on the literature review, preliminary analysis of A/R crash rates, and consultation with NHTSA, the final high-risk groups were selected. Six

high-risk groups were selected along with a one-tenth sample of the general N.C. driving population to be studied throughout the remaining steps of the project. The final six high-risk groups selected for further analysis were:

1. Young males, 16-20 years old
2. Young males, 21-24 years old
3. Persons previously convicted of DUI
4. Persons with three or more moving violations
5. Persons recently divorced
6. Persons recently released from prison

At this stage two separate files were developed for each group--one containing two-thirds of the group and the other containing the remaining third, identified by taking every third subject. The two-thirds sample was used in developing the models and the remaining third was reserved to conduct concurrent validity tests of the models once they were completed.

2.3 Model Development

Once the high-risk groups had been identified, it was necessary to select the appropriate multivariate technique for developing the predictive models. The models were developed using driver-related information which was known on or before December 31, 1974 to predict probabilities of A/R crash involvement in 1975.

Since both the dependent variable (presence or absence of an A/R crash in 1975) and most of the independent variables (such as sex, violation types, accident types, and the like) were of a categorical nature, it was decided to use a modelling technique specially developed for categorical data. The GENCAT technique (Grizzle, Starmer, and Koch, 1969) was selected and used in developing

a separate model for each of the six high-risk groups and for the general population sample. The basic modelling steps are outlined below and are described in more detail in Chapter 5.

2.3.1. Univariate variable selection.

Each of several demographic, accident, and driver history variables were examined for each group to determine how they could best be used to account for the group's variation in 1975 A/R crash involvement. For many of the variables, it was necessary to select the optimum levels or value ranges for the variable in accounting for A/R crash variation. An example is the driver history variable "days under suspension or revocation." The data were examined to determine the optimum way to group the values of that variable as in, for example, (0, 1 or more) or (0, 1-30, 31 or more), etc.

For driver history variables an optimum time frame for accumulating values for each variable was also determined. Ranges examined included time periods of from six months prior to December 31, 1974 to up to four years prior.

2.3.2. Multivariate variable selection.

After the optimum levels and time frames were selected for the variables for each group, the variables to be included in the actual model fitting step were selected in a stepwise manner. The steps were, first, select that variable which accounted for the most variation in 1975 A/R crash involvement: then select the variable which, in combination with the first, accounted for the most additional variation, and so on until no more significant variables remained or the cell size became too small to be practical.

2.3.3. Model development and fitting.

After the predictor variables were selected for each group, categorical data models were developed to predict 1975 A/R crash rates for each group.

These models delineate several subgroups within each high-risk group and assign predicted proportions of each subgroup expected to be involved in an A/R crash in a year. Thus, for each group, a set of proportions is provided which range from well below the total average risk for some subgroups to well above that average for others.

2.4 Validity Testing

2.4.1. Concurrent validity testing.

As mentioned in 2.2.6. above, one-third of each group was reserved to conduct concurrent validity tests on the final models. Goodness of fit statistics were computed for each group comparing the predicted proportions developed on the basis of two-thirds of the group with the actual proportions which experienced A/R crashes in the one-third sample.

2.4.2. Prospective validity testing.

A truer test of the models' predictive performance is to examine how well the models predict A/R crash performance in a subsequent year. This issue was addressed in two ways.

2.4.2.1. Original groups - 1976 crashes.

The actual A/R crash performance in 1976 of each of the groups, as identified by data available as of December 31, 1974, was examined. Thus, although the models were designed to predict one year ahead, it was decided to examine their two-years-ahead predictive ability as well. Goodness of fit statistics comparing predicted versus actual A/R crash experience were computed for each group. Rank correlation of the subgroups within each group was also examined.

2.4.2.2. Newly identified groups - 1976 crashes.

The most appropriate test of the models' predictive ability as designed was made by identifying new groups using data available through 1975 and examining their A/R crash experience in 1976 as compared to that predicted

by the models. The same tests outlined in 2.4.2.1. were conducted on these new groups.

2.5 Development of User Manual

A major objective of the project was to present the models in a framework in which they could be applied by alcohol or motor vehicle program administrators. To that end a joint and parallel effort to the model development was made in developing a user manual which would provide tools for countermeasure program selection, implementation, and evaluation.

2.5.1. Review countermeasures.

The traffic safety literature was reviewed to identify potential countermeasure programs which might be appropriate to the high-risk drivers. To assist in countermeasure selection, an attempt was also made to extract from the evaluation literature expected levels of effectiveness for various countermeasures.

2.5.2. Select economic analysis technique.

An appropriate economic analysis technique was selected to be presented as an aid in prioritizing potential countermeasure programs on a cost-effectiveness basis. This procedure was computerized for subsequent use.

2.5.3. Review evaluation literature.

The general evaluation literature was reviewed in order to select appropriate evaluation designs and procedures to guide in the implementation and evaluation of any countermeasure activities which might be initiated using the models.

2.5.4. Develop user manual.

The results of the steps outlined in 2.5.1.-2.5.3. were used along with the predictive models developed during the project to construct a user's

manual which may be used to assist in identification of individuals in need of countermeasure activity, selecting countermeasures on a cost-effectiveness basis, and implementing such countermeasures in a way that their true effectiveness in terms of A/R crash reduction can be evaluated.

The remainder of this Volume and Volume II, the User's Manual, outlines in more detail the procedures followed and results obtained in pursuit of these project objectives.

CHAPTER 3 - HIGH-RISK GROUP SELECTION

3.1 Introduction

This chapter presents for each high-risk group the results of the initial steps taken in identifying high-risk groups for the subsequent modelling and validation procedures described in Chapters 5 and 6. The two basic steps were first, to review the literature to identify groups known to be at a high risk of A/R crash involvement, and second, conduct a preliminary data analysis on those groups to determine if they actually did have a high A/R crash involvement rate in N.C. in 1973-1975.

As mentioned in 2.2.2, few studies have been conducted which specifically address the risk of A/R crash involvement for particular segments of the driving population. So, in many cases, the review presented here will report separate studies which evidence high alcohol consumption on one hand and high crash involvement on the other hand. Thus, in the case of some groups it was necessary to assume that the two would be likely to occur simultaneously.

To confirm the results of the literature review, preliminary data analyses were done before multivariate modelling procedures were begun. For each of the high-risk groups selected on the basis of the literature survey and consultation with NHTSA, a further analysis of North Carolina crash data was conducted before it was selected for modelling. Two measures of the appropriateness of the groups were taken. One was the ratio of the population A/R crash rate of the high-risk group to the general population's A/R crash rate. This will be called the risk index. The other was the proportion of all N.C. A/R crashes that the high-risk group accounted for. This will be called the impact index.

These indices were computed for the years 1973, 1974 and 1975 for each of the candidate high-risk groups as they would be identified at the end of 1975.

Thus, for example, if an individual did not meet the criteria for group inclusion until the end of 1975, he was still considered, for the purposes of these preliminary analyses, a member of the high-risk group in 1974 and 1973 as well. This could affect the risk and impact indices for some groups for these earlier years. For example, a male who became 16 near the end of 1975 would be in the young male group yet unlikely to be involved in A/R crashes in 1973 and 1974. For the modelling stages of the project this problem does not exist in that groups were identified as of December 1, 1974 and A/R crashes were examined in 1975.

The risk and impact indices for each high-risk group selected through the literature review are presented at the end of the discussion of the literature for that group.

In this chapter the rationale for the inclusion of six high-risk groups is presented. That does not mean that other groups were not considered or that still others might not warrant further study along these lines. However, through the review it became apparent that these particular groups would be most appropriate for this study. For each candidate group, issues such as the ease of identification by alcohol program administrators, the ability to establish absolute descriptive criteria, the potential impact on the total A/R crash problem, the potential of having particularly high-risk subgroups, and the ability to address special issues, such as transient situational stress and the like were all considered throughout the review. Thus, such groups as problem drinkers, older drivers, women drinkers, etc., were all carefully considered, but for one or more of the reasons above they were not selected in favor of the six high-risk groups.

The high-risk groups which were selected for further study were: young males, 16-20; young males, 21-24; persons previously convicted of driving under

the influence; persons with three or more moving violations; persons recently divorced; and persons recently released from prison. Additionally, a one-tenth sample of the general driving population was selected for modelling and validation in order to assess the benefit achieved, if any, from preselecting high-risk groups.

The rationale for the selection of the high-risk groups follows.

3.2 Young Males, 16-20 and 21-24

There is a consensus in the highway safety community that young males are at an elevated risk of traffic accident involvement (Waller, 1971). Many factors are cited as contributing to this over-involvement. They include inexperience in driving (Goldstein, 1971), a predisposition towards exhibiting risk-taking behavior (Waller, 1971), high exposure to high-risk driving situations and times (Pelz and Schuman, 1971), and inexperience in alcohol use, which may lower the levels at which alcohol affects driving performance (O'Day, 1970).

3.2.1 Young males and crashes.

Young males have been found to be overrepresented in A/R crashes. Preusser, Oates and Orban (1975) reported on interviews of a sample of male New York drivers aged 16-24 and 35-49. They found that 14 percent of the young drivers vs 5 percent of the older drivers reported having an A/R crash within the previous three years. They also reported on the distribution of fatally injured drivers in Nassau County, New York, 1967-1971 by age and alcohol. Thirteen percent of those showing positive BAC readings were 19 or younger and 22 percent were 20-24, while the proportion of licensed male drivers accounted for by these two age groups were 7 and 12 percent

respectively. Ninety-three percent of all drinking, fatally-injured drivers were male.

Rosenberg, Laessig and Rawlings (1974) reported on 1968-1971 fatally injured Wisconsin drivers, excluding Milwaukee, for which blood alcohol determinations were made. They found that their sample was predominantly young (the 16-19, and 20-24 age groupings were the two largest), and that 48.3 percent and 70.2 percent of the fatally injured drivers were at BAC's $\geq$ .10 in these two age groups. Those aged 16-25 constituted 45 percent of the study group while accounting for 23 percent of all male licensed drivers in Wisconsin.

The Minnesota Department of Public Safety (1970) reported that among 1969 Minnesota driver fatalities 16-24, over 60 percent had positive BAC readings. Ninety-three percent of all alcohol-involved driver fatalities were male. Perrine, Waller and Harris (1971) in a study including drivers fatally injured in Vermont between July 1, 1967 and April 30, 1968 reported 60 percent of the fatally injured drivers 24 and younger had positive blood alcohol readings.

In an analysis of the age distribution of fatally injured drivers in Nassau County, New York, from 1968 through June 1973 (85 percent of which were male), Ulmer and Preusser (1973) found the distribution to be "significantly different (x^2 = 31.046, d.f. = 11, $p < .01$) in the direction that drinking drivers killed tend to be younger than fatally injured drivers who had not been drinking."

3.2.2 Young males and drinking.

Marden and Kolodner (no date) reviewed studies of alcohol use among adolescents. Survey results of males 16, 17 and 18 from 1970-1974 consistently showed that in excess of 75 percent drank. Rachel, et al.

(1975), conducted a national probability sample survey of youth aged 13-18 and reported that nearly 40 percent were moderate to heavy drinkers.

Cahalan (1970) reports in Problem Drinkers, a study based on a national probability sample of adults in the continental U.S., that "among men, the prevalence of (drinking related) problems (in the aggregate) is highest among those in their twenties." Cahalan and Room (1974) in a further analysis of males from an enlarged data set stated "younger men (especially those aged 21-24) have the highest rates of both very heavy and steady fairly heavy drinking."

3.2.3 Young males, age groupings.

Many researchers (Carlson, 1972; Goldstein, 1971; O'Day, 1970; and Zylman, 1973) have theorized that there may be two factors which make major contributions to the young male's overinvolvement in crashes. They feel that inexperience in driving coupled with inexperience in drinking makes young males particularly vulnerable to A/R crashes. Some (O'Day, 1970) think that the high rate for the younger half of the group may be attributable to driving inexperience and that for the older half due to drinking inexperience.

There is a good deal of variation between states in the age at which it is legal to purchase different categories of alcohol beverages. (In N.C., beer and wine can be purchased legally by 18 year-olds while distilled liquor cannot be bought until age 21.) Thus, different types of drinking behavior are likely to occur during the age span of 16-24. Since widely different drinking and driving behavior patterns could occur across this wider age span the group was split into 16-20 and 21-24 in order to describe more homogeneous groups for further analysis.

3.2.4 Young males, impact and risk indices.

The risk and impact indices as defined in 2.2.5 and 3.1 for the two young male groups are presented in Tables 3.1 and 3.2.

Table 3.1 A/R crash risk indices 16-20 and 21-24 year old males by year. (Risk index for average driver is 1.0.)

	Year		
Group	1973	1974	1975
Males, 16-20	1.25	2.02	2.72
Males, 21-24	2.65	2.74	2.62

Table 3.2 A/R impact indices 16-20 and 21-24 year old males by year.

	Year		
Group	1973	1974	1975
Males, 16-20	5.52	11.12	17.31
Males, 21-24	17.88	18.23	17.12

Examination of Table 3.1 indicates a heightened driver population based risk for both groups with a more consistently high one for the older group. Likewise, the older group's impact on the total A/R crash problem is also consistently nearly one-fifth while the younger group builds to that level. It should be remembered that for the purposes of this preliminary analysis the groups were defined as of the end of 1975. Thus, for the younger group, fewer drivers were in the sample for 1974 and 1973, and thus, fewer crashes would be expected. The 1975 figure is probably most representative of the groups' actual performance.

3.3 Persons Convicted of Driving Under the Influence of Alcohol (DUI)

Persons who fall into this group have already exhibited one of the behaviors necessary to an A/R crash. Of course, this drinking behavior while driving also indicates an increased risk of crash involvement.

In most jurisdictions in the U.S., a BAC of .10 provides probable cause for conviction of DUI. According to Borkenstein (1964), BAC's of .10 were associated with an increased risk of causing a crash that was seven times as high as the risk for drivers with no alcohol. Perrine, et al. (1971) reported a relative probability of having a fatal crash of around 8 for BAC's of .10 to .12, when compared to a reading of .00.

The average BAC level for persons arrested for DUI reported in many jurisdictions ranges from .17 to .20. The authors above reported a crash risk of 25 to 1 for drivers at the .15 level compared to a non-positive reading. So the level exhibited by the average DUI arrestee places him at an even higher risk.

Recidivism rates in many ASAP projects were high among DUI arrestees (U.S. DOT, 1974), meaning that after being arrested for DUI they subsequently exhibited and were arrested for the same driving behavior which again put them at a high risk of A/R crash involvement.

Filkins, et al. (1970) conducted a case history investigation of 616 Wayne County traffic fatalities from July 1967 through August 1969. They found a significant relationship ($p = .02$) between previous DUI convictions and blood alcohol level. They also compared a sample of DUI offenders with the general driving population and found them to have nearly three times as many accidents. Thus there is a good deal of evidence that DUI offenders are a heightened risk of crash involvement.

3.3.1 DUI, impact and risk indices. The risk and impact indices for the DUI group are presented in Tables 3.3 and 3.4.

Table 3.3 A/R crash risk indices for DUI group by year.

Year		
1973	1974	1975
12.31	12.12	10.10

Table 3.4 A/R crash impact indices for DUI group by year.

Year		
1973	1974	1975
34.71	33.91	27.61

As inspection of Tables 3.3 and 3.4 reveals, the DUI group is at an extremely high risk of A/R crash involvement and accounts for nearly a third of all A/R crashes.

3.4 Persons with Three or More Moving Violations (3+ Violations)

Traffic violations have historically been found to be one of the best correlates of accident involvement. That certain driving acts are considered violations of traffic law is predicated on the assumption that the act is unsafe and more likely to result in a crash than other "more normal" types of driving behavior. Thus persons who are repeatedly cited for moving traffic violations should be more likely to experience a crash than those who are not. This logic is confirmed in the literature.

Williams (1958) examined the driving records of 95,000 California drivers and for a three-year period found a correlation of .26 between violation convictions and accidents. As number of violations increased, mean number of accidents tended to increase. For the group of drivers with nine or more

violations in a three-year period the mean number of accidents was over six times that of the zero violation group.

Burg (1968) examined the six-year driving records of a sample of California driver license applicants. In a concurrent three-year period he found a number of convictions to be correlated with number of accidents (.300). Examining convictions in one three-year period in relation to accidents in the next three years, he found somewhat lower correlations (.152).

Other researchers have also found convictions to be significantly correlated with subsequent accident involvement. Peck, McBride and Coppin (1971) examined the records of 148,000 California drivers and found a significant correlation between violation convictions in one year and accidents in the next (a range from .057 to .089) ($p < .01$). Marsh and Hubert (1974) examined the driving record of 6795 male negligent drivers and found subsequent accident experience to be significantly associated with hazardous driving violation convictions ($r = .021$; $p < .10$).

Filkins et al. (1970) reported on previous driving violations of a group of fatally injured Michigan drivers. They found previous convictions to be significantly associated with BAL among those drivers ($p = .006$).

3.4.1 3+ Violations group, risk and impact indices

The risk and impact indices for persons with three or more moving violations are presented below.

Table 3.5. A/R crash risk indices for 3+ violations group by year.

Year		
1973	1974	1975
6.45	6.60	5.91

Table 3.6 A/R crash impact indices for 3+ violations group by year.

Year		
1973	1974	1975
41.12	41.71	36.62

This group, as measured by the risk index, has six times the A/R crash rate of the general population and accounts for over one-third of N.C. A/R crashes while making up only about seven percent of the driving population.

3.5 Persons Recently Divorced

Consideration of divorce as a possible A/R crash predictor variable is based on the premise that stressful life events may tend to make certain individuals more likely to become accident-involved. Several accident studies which have considered marital status as a variable have revealed heightened accident involvement for divorced persons in both the context of A/R crashes and crashes taken as a whole. Additionally some studies of divorce have indicated alcohol as a factor in precipitating the divorce itself.

Borkenstein et al.'s (1974) classic study of A/R crashes in Grand Rapids showed divorced persons to be overrepresented in A/R crashes as compared to their site and time-matched control group.

The Institute for Research in Public Safety (1973) compared accident - involved drivers with a sample from the general driving population. They computed an involvement ratio for several descriptor variables of the drivers which compares their proportion in the accident sample to their proportion in the control sample. For the marital status variable, divorced persons were the

most overinvolved group with an accident involvement ratio of 4.7. For alcohol-related crashes, the involvement ratio was 9.1.

A study of divorce and accident involvement was conducted by McMurray (1968). She compared a sample of Washington State drivers in the process of divorce proceedings with an age and sex adjusted control group from the general driving population. The persons involved in divorce proceedings had from 42.70 percent to 81.78 percent more accidents than the control group depending on sex and role in the litigation (plaintiff or defendant). The total number of accidents and violations per individual was 104.16 percent higher in the divorce group than in the control group.

Carlson (1973) examined the relationship of BAC distribution to marital status for drivers stopped in a nighttime roadblock survey of Washtenaw County, Michigan. He found the divorced and separated classification to be significantly associated with increased BAC levels ($p < .01$).

Filkins et al. (1975) compared marital status of the U.S. male population 18 and older, the National Roadside Survey (NRS), and accident cases from the Collision Performance and Injury Report (CPIR) file. They found that 4.5 percent of the U.S. population file were divorced or separated persons and that 13.9 percent and 15.3 percent of the NRS and CPIR samples, respectively, were divorced persons who had driven with BACs of .10 or greater.

In a study of 600 couples seeking divorce in Cuyahoga County, Ohio, Levinger (1966) reported on the complaints aired in mandatory joint counselling interviews. He reported that 26.5 percent of the women indicated that excessive drinking on the part of the husband was one of the sources of their marital disharmony. Kephart (1954) studied a 25 percent random sample of common pleas court records of 1434 divorces in Philadelphia between 1937 and 1950. Drinking was reported by the plaintiff as an alleged causal factor in 21.1 percent of the

cases. In fact, excessive drinking, although not legal grounds for divorce, was reported "more frequently than any other single factor except desertion and indignities, both of which are legal grounds for divorce."

Thus there is ample evidence in the literature to indicate that divorced persons are at increased risk of crashes, alcohol problems, and alcohol related crashes.

3.5.1 Divorce group, risk and impact indices.

To determine if divorced persons were overinvolved in North Carolina A/R crashes, a listing of persons granted divorces in North Carolina in 1974 was obtained from the Department of Human Resources. The accident records of those persons who could be matched with Division of Motor Vehicles records were queried and compared with those of the general driving population. Tables 3.7 and 3.8 reflect the resulting A/R crash risk and impact indexes for the divorce group.

Table 3.7. A/R crash risk indices for divorce group by year.

	Year	
1973	1974	1975
2.88	2.65	1.86

Table 3.8. A/R crash impact indices for divorce group by year.

	Year	
1973	1974	1975
1.20	1.10	0.75

From these tables it becomes apparent that the highest risk year for this group of persons divorced in 1974 is the year before the divorce is

granted. This is in line with the theory of some observers that the most stressful period for persons terminating a marriage is the year that the separation and divorce proceedings are underway. However, the official records of divorces are not available on a statewide basis until final divorce is granted, and thus would not be available for predicting until the year after final divorce. Though the crash experience in that year is less extreme than in the preceding two years, it is still nearly two times that of the general population. Since the total divorce group constitutes only a small portion of the total driving population its A/R crash impact index represents only about one percent of all A/R crashes even though they have them at about twice the rate of the general driving population.

3.6 Persons Recently Released from Prison

Another group undergoing a stressful period in their lives is persons recently released from prison. There is evidence in the literature that this stress is also related to crashes.

Harano (1974) developed predictive models for both collision involvement and traffic offense convictions. He constructed models using driver record, criminal record, questionnaire, and psychometric test data on one group of 430 drivers and, using only driver record and criminal record, on an enlarged group of 1196. He split each sample into two groups--one to construct the models using multiple stepwise regression and one on which to cross-validate the models. For the 430-individual group, seven variables entered the construct equation at the $p < .01$ level for predicting collisions. One of these variables was from the criminal record data. However, cross-validation indicated a non-significant cross-validity coefficient of .03. On the larger data set, the

construct equation contained age and criminal violations as the only predictors significant at the $p < .01$ level. For this group a significant cross-validity coefficient of .11 ($p < .01$) was obtained.

Harano, McBride and Peck (1973) considered some criminal information variables among 393 variables examined for 427 male drivers to identify ones of use in equations predictive of accident involvement. Though none of the criminal variables entered the predictive equations, burglary/robbery arrests and the category "other type arrests" had F values of 2.18 and 4.84, respectively.

Pollack et al. (1972) reported that they were able to develop a model contrasting drinking drivers and non-drinking drivers among a sample of nearly 4000 drivers. From extensive data accumulated on their subjects, five variables were selected for the model. Non-traffic arrests was one of the five predictor variables.

Li and Waller (1976) examined the one year prospective driving record of persons referred to the N.C. habitual offender program. They found that 15 percent of persons with a prison record incurred alcohol related driving offenses in that year compared to seven percent of those with no prison record.

Examination of the N.C. Department of Correction Statistical Abstract for 1974 reveals that a large number of persons in prison are there as a result of driving or drinking related offenses. Of 12802 persons admitted to the North Carolina prison system in 1974, 1238 received their sentences as a result of conviction of DUI offenses, 800 more were admitted for other traffic offenses and an additional 304 were admitted for habitual or public drunkenness. Additionally, it is known that a large number of the crimes which result in incarceration occur after alcohol consumption. For example, in a Law Enforcement Assistance Administration of 8711 male prison inmates, 43 percent

reported having been drinking at the time of the crime (U.S. Department of Justice, 1975).

That excess drinking patterns might be found in a group of persons released from prison is supported by the findings of Guze, et al., (1962) who studied a series of 223 criminals recently or soon to be released from prison. They found through psychiatric evaluation that 43 percent of their subjects exhibited alcoholism.

3.6.1 Prison group, risk and impact indices.

A listing of persons released from prison in N.C. in 1972 was obtained and matched with the driver history and accident files in order to access the involvement of this group in A/R crashes. Tables 3.9 and 3.10 show the A/R risk and impact indices for this prison group.

Table 3.9. A/R crash risk indices for prison group by year.

Year		
1973	1974	1975
6.29	5.52	6.29

Table 3.10. A/R crash impact indices for prison group by year.

Year		
1973	1974	1975
.91	.77	.83

Examination of the tables reveals that the prison group has a high A/R crash rate, but, because it is small, it does not account for a large proportion of all A/R crashes.

3.7 Final Group Selection.

After consultation with NHTSA, it was decided to retain all six high-risk groups for the modelling phase of the project. The first four groups, young males (16-20), males (21-24), persons convicted of DUI and persons with three or more moving violations were selected because of both their high A/R crash rates and the large proportion of all A/R crashes they account for. The last two groups were selected because of their high A/R crash rate and because they represented groups likely to be undergoing transient situational stress.

The groups are not mutually exclusive. For example, a 22 year old male could also be divorced, recently out of prison, have a DUI and so on. Thus, the proportion of all A/R crashes accounted for by the individuals who make up the six high-risk groups is somewhat less than the sum of impact indexes for all groups. The crash record of each individual that met the basic criteria for high-risk group inclusion was examined once, whether he appeared in more than one group or not, to determine the proportion of all A/R crashes accounted for by the high-risk individuals. They accounted for approximately two-thirds of A/R crashes. The percentages for 1973-1975 appear in Table 3.11.

Table 3.11. Percent of all A/R crashes accounted for by all individuals in high-risk groups by year.

Year		
1973	1974	1975
65.63	68.21	66.55

Thus, a large proportion of the A/R crash problem was found to be attributable to the six high-risk groups selected.

In the following chapter the data base used in the modelling and model testing process is described.

CHAPTER 4 - DATA BASE

4.1 Introduction

This chapter contains a description of the data base used in identifying the high-risk groups, developing the predictive models and testing their predictive validity. As was noted earlier, statewide data from North Carolina were used. The underlying rationale for use of this data base is as follows: first, because of a need for a large data base to identify extremely high-risk subgroups of a large enough size for meaningful countermeasure programs, the most likely application of the predictive models is on a statewide basis or in large metropolitan areas; second, individual driver records and accident files are more routinely collected on a statewide basis than in smaller governmental jurisdictions; third, multi-state data sources were not felt to be as desirable because of a lower likelihood of compatible records-keeping systems and multistate driver-related countermeasure activities; and finally, North Carolina was selected because of its fundamentally sound computerized driver history and accident data collection system, because an active program exists in the state to train enforcement officers in accident reporting and because its accidents are reported according to uniform criteria and on the same form statewide at all jurisdictional levels.

Because a major project objective was that the modelling procedure used and the models developed from North Carolina data be suitable for use in other states which may wish to use them, a basic description of North Carolina in areas relevant to this study is presented in the initial section of this chapter for comparison with other states. The individual data bases used in this study are then described as well as the data merging operations required to obtain the final study record. Finally the number of records used at critical steps in the data reduction process are presented.

4.2 North Carolina Data Characteristics

North Carolina is a predominately rural state with major metropolitan areas located in the central and western areas. Population in these major urban areas ranges from 100,000 to 300,000. Rural road mileage in the state is almost totally under the centralized control of the state Division of Highways, and because the emphasis placed in the past on upgrading farm to market roadways, there are many miles of paved, two-lane rural road. Of the 71,000 miles of state roadway, over 80 percent is classifed as secondary roadway. The rural primary mileage is approximately 88 percent two-lane and 12 percent four-lane divided. There are 709 miles of rural Interstate and 162 miles of urban Interstate highway in the state.

In January of 1977, there were approximately 3,400,000 licensed drivers in North Carolina. To obtain a valid license, an applicant must be at least 16 years old and must pass knowledge, vision, signs and road tests. License renewal is required every four years, at which time signs, vision, and, in some cases, knowledge and road tests are again given. The percentage of licensed drivers by age is presented in Table 4.1.

Table 4.1. Number and percentage of licensed drivers by age and sex.

	Sex		
Age	Male (Col. %)	Female (Col. %)	Total (Col. %)
16-21	282,453 (15.64)	246,526 (15.44)	528,979 (15.55)
22-25	209,419 (11.59)	191,741 (12.01)	401,160 (11.79)
26-30	238,698 (13.21)	223,706 (14.01)	462,404 (13.59)
31-45	468,109 (25.91)	444,716 (27.86)	912,825 (26.83)
46-54	240,099 (13.29)	216,382 (13.55)	456,481 (13.41)
55+	367,697 (20.35)	273,267 (17.12)	640,964 (18.84)
Total	1,806,475 (53.09)	1,596,338 (46.91)	3,402,813

All accidents which involved personal injury and/or $200 property damage are investigated by either city or county police or the N.C. State Highway Patrol, and all are reported to the N.C. Division of Motor Vehicles using a uniform report. In 1977, the total number of accidents was 145,670 with 53.5 percent occurring in rural areas and 46.5 percent occurring in urban areas. The State Highway Patrol investigates approximately 47 percent of all crashes.

The 1977 accident total included 1261 fatal accidents (0.87%), 51,264 nonfatal injury accidents (35.19%) and 93,145 property damage only accidents (63.94%). In the fatal accidents, 1437 deaths occurred, with 85.5 percent occurring in rural areas and 14.5 percent in urban areas.

When the data used in this study were collected North Carolina's alcoholic beverage control laws were rather conservative, with distilled liquor only being sold at state controlled stores in counties or cities which had passed referenda establishing such stores.[1] Beer and wine could be sold by licensed private businesses where local referenda had been approved. In 1973, 13 of the 100 counties remained completely dry (i.e., no beer, wine, or liquor). Of the remaining counties, 30 are semi-dry counties where certain local municipalities have established either liquor, beer, and/or wine sales. Perhaps because of these rather restrictive laws, average per capita consumption in North Carolina as computed by legal sales receipts was approximately 20 percent below the national average. As has been noted by some researchers, this low average may be somewhat misleading in that, being based on consumption of legal alcoholic

---

[1] In 1978, a local option liquor by the drink bill was passed by the State Legislature. This bill allows local jurisdictions that already have liquor stores to vote on liquor by the drink. The first such vote occurred in September, 1978. However, for model development purposes, the above description is more relevant.

beverages, it does not take into account the supposedly thriving bootleg liquor industry in the state.

However, even assuming these somewhat conservative consumption patterns, the problem of alcohol and driving remains. In 1975 accidents in which the drinking status could be determined, alcohol-involvement (not necessarily impairment) was noted by the investigating officer for 11.6 percent of all drivers in rural accidents and 6.0 percent of drivers in urban accidents. For fatal accidents the corresponding figures were 25.9 percent for drivers in rural accidents and 19.8 percent for drivers in urban accidents. Data collected by the North Carolina Office of the Chief Medical Examiner on BACs of fatally injured single vehicle operators reflect over 50 percent with a BAC $\geq$ .10.

As noted above, rural accident investigation and traffic law enforcement is the primary mission of the N.C. State Highway Patrol. In 1975, the 1,162 man patrol issued 480,585 traffic citations of which 35,911 were for first offender DUI's and 5,062 were for the second or subsequent DUI violation. With a strong statewide breath testing program, an implied consent statute, and a *per se* law, the conviction rate for DUI arrests has been between 62-63 percent over the past three years.

4.3. Data Files

Information from four different files was used in the project. The files were the driver history file, the accident file, the divorce file, and the prison file.

4.3.1. Driver history file.

This file consists of approximately 3.8 million variable-length records containing the driver history of each licensed North Carolina driver and some drivers to whom a valid license has not been issued but who have come to the

attention of the authorities due to a violation or accident. Each subject's file contains basic information on age, race, sex, and the initial and most recent licensing activity. As additional activity relevant to driver licensing and regulation occurs for an individual, it is added to the record resulting in the variable length format. Examples of this activity are violations, accidents, convictions, warning letters, and suspensions or revocations.

For persons arrested for DUI from January 1, 1974 on, a separate additional confidential trailer has been added to their record containing such information as BAC, time of arrest, and disposition, regardless of whether the person was convicted of the offense of DUI. This trailer is called the RATERS trailer. Thus, information about alcohol use, which normally would no longer be available, is retained for persons convicted of a lesser included offense such as reckless driving or for persons acquitted or nol prossed.

The driver history file is arranged by driver license number. For the model development phase of this current project, information available through December 31, 1974 was used. For the prospective predictive validity testing phase, information available through December 31, 1975 was used.

4.3.2. Accident file.

This file contains detailed descriptive information on accidents reported in North Carolina. Information such as driver name and license number, accident type, crash severity, injury severity, time of day, weather conditions, and alcohol involvement appears on each record in this file. Approximately 140,000 accidents with descriptive information on 250,000 vehicles and their occupants are recorded each year on this file. For the model development phase of this project, 1973 and 1974 accident information was considered for use as independent variables; and alcohol-related crashes

in 1975 were used as the criterion or dependent variable. In the prospective validity testing phase, 1974 and 1975 accidents were used as predictive variables and 1976 A/R crashes became the criterion variable.

4.3.3. Divorce file.

This file, obtained from the N.C. Department of Human Resources, is an alphabetic listing of persons granted a divorce in North Carolina. The information on this file includes name, race, and county of residence. Each year the file contains approximately 40,000 names. For the modeling phase, a file of persons granted a divorce in 1974 was used. For the prospective validity phase, 1975 data were used.

4.3.4. Prison file.

This file, obtained from the N.C. Department of Correction, contains information on persons released from prison. Identifying information such as name, age, race, sex and former address were extracted for use in the study. Approximately 10,000 persons are released each year of which approximately 6,500 have not been returned to prison two years later. For the modeling phase persons released in 1972 and not returned to prison by the end of 1974 were considered. For the prospective validity testing phase, persons released in 1973 and not returned to prison by the end of 1975 were considered.

4.4. File Merging and Study Record Development

To perform the modelling steps, relevant information from each of the four files used had to be merged into one file combining all of the information for each individual into a single record. In order to perform the merge, a single identifying variable common to all files had to be assigned to each individual. This variable was North Carolina driver license number. That number appears on

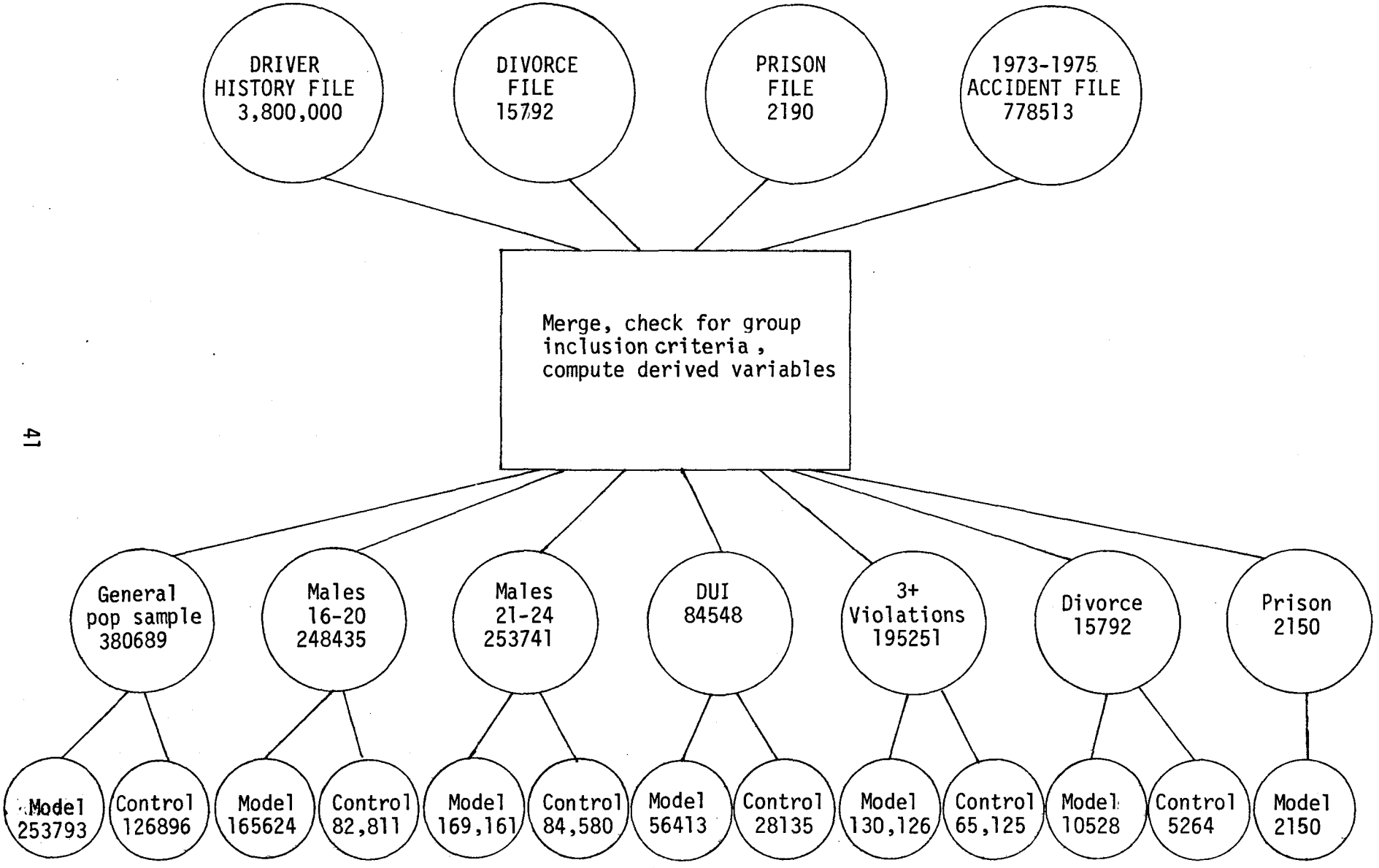
the driver history file and, for N.C. drivers, on the Accident File. The driver license number does not appear on either the prison file or the divorce file.

For the prison and divorce files, a file was prepared containing data elements common to both that file and the driver history file. For the prison file, those data elements were name, age, race, sex, and former address. For the divorce file, they were name, age, and county of residence. These files were sent to the Division of Motor Vehicles where a computer matching routine was applied to match them with the driver history file. In some cases multiple matches were obtained. These cases were manually inspected in order to identify the true match if possible. If the correct match was not identifiable, the individual was not carried further in the study. Because of these non-matches, considerable shrinkage in group size occurred for the divorce and prison groups in the merging process. For the 1974 divorce group 40,098 names were sent to DMV, and 15,752 driver license numbers were obtained. For persons released from prison in 1972, a file containing information on 7,113 persons yielded 2,190 driver license numbers.

In the next steps, the four files were merged together, the single consolidated file read, records meeting the criteria for group inclusion identified, and seven new files (one for each group) developed. With the exception of the smaller prison file, each group was then subdivided into a subgroup for model development (2/3) and a subgroup for concurrent validity testing (1/3). Figure 4.1 depicts the sequence described above.

As part of the processing of the records, certain computed variables were created from the data. These included tallies of the total number of crashes, A/R crashes, and night crashes 1973-1975, and total crashes and A/R crashes in 1973-1974. Detailed information on the most recent 1973-1974 crashes (up to a total of three) by each individual was also retained.

Figure 4.1. File Merge and Group Identification



Violation information from the driver history file was converted from variable length format to a fixed length format by tabulating various types of violations over six-month periods beginning December 31, 1974 and working back in time for up to eight such periods (four years). A similar procedure was followed for a one-year period for alcohol arrest data from the RATER's trailer in the driver history file. Appendix A, Alcohol Model Study Record Format, presents the full format of the study record.

4.5. Group Size Reduction During Analysis

As mentioned in section 2.3 of the Methodology Chapter and discussed in more detail in Chapter 5, variables for use in the predictive model development were selected from the variables in the Study Record Format based on their association with subsequent A/R crashes.

For each of the high-risk groups separately and for the sample of the general driving population, each of the variables appearing on the format in Appendix A was considered as a possible predictor variable. For the driver history variables, which were available in varying lengths of time by combining six-month periods, several lengths of time were considered in the selection of predictor variables. If a particular variable was highly correlated with A/R crashes and was selected for inclusion in the modeling process but did not appear on an individual's record, then that individual record would no longer be available for consideration in the model. In particular, for driver history variables where combinations of six-month periods, say a two-year or a three-year period, were used to define predictor variables, those individuals with driver records of a shorter duration would not be included in the modelling process. This led to considerable shrinkage in group size due to unavailable data. In Table 4.2 the group size for each group is shown before variable selection and

after variable selection. These numbers reflected the two-thirds of the group which was used in the model development.

Table 4.2. Modeling group size before and after predictor variable selection.

Group	Number in group before variable selection	Number in group with all variables selected for modeling
General population sample	253,793	177,239
Males, 16-20	165,624	91,938
Males, 21-24	169,161	68,306
DUI	56,413	38,657
3+ Violations	130,126	125,850
Divorce	10,528	8,625
Prison	2,190	1,989

One can see from inspection of Table 4.2 that considerable shrinkage in group size occurred during the variable selection process. This has implications for both countermeasure selection and potential impact of countermeasures on the total A/R crash problem. These issues are discussed in more detail in Chapter 7. In the following chapter the actual steps taken to develop the predictive models are described and the predictive models themselves are presented.

CHAPTER 5 - MODEL DEVELOPMENT

5.1 Introduction

This chapter describes the various steps that were followed to derive models for predicting future alcohol-related crashes. These steps involved:

a. The choice of statistical methods,

b. The choice of the most appropriate time frame and the levels for each potential explanatory variable,

c. The selection of the most important variables for inclusion in the model, and

d. The fitting of the models to the data.

In the final section of this chapter, the resulting models are presented.

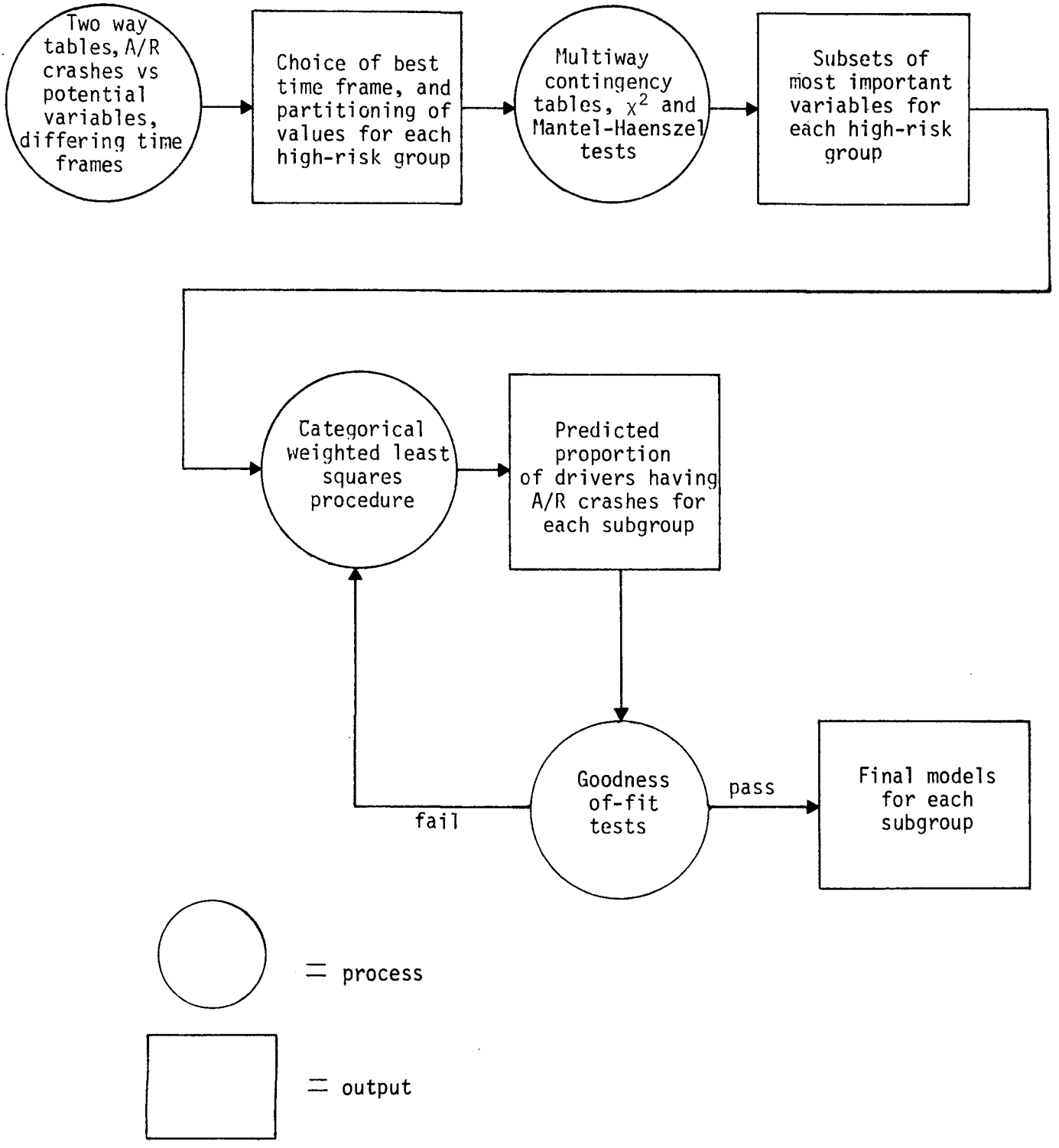
Figure 5.1 illustrates the sequence of steps in developing the models. This procedure was followed for each of the high-risk groups and for a sample of the general driving population.

5.2 Choice of Statistical Methods

The basic problem which had to be solved was that of relating alcohol crash involvement to certain characteristics of drivers and their past driving histories. Several statistical methods are available for dealing with such problems.

Perhaps the most widely used methods are stepwise multiple regression analysis and stepwise discriminant analysis. Computer programs are readily available for both of these methods, which solve both the variable selection problem and the model fitting problem simultaneously. The test statistics for both methods are based upon assumptions of normality (of the dependent variables for regression analysis, and of the independent variables in discriminant analysis), while in the present application both the dependent variable (number

Figure 5.1 Flow Chart - Model development procedure.



of alcohol-related crashes) and nearly all of the potential independent variables are discrete-valued. Thus, while either regression analysis or discriminant analysis (with suitably defined dummy variables) could be used in the development of A/R crash predictor models, analogous methods specifically designed to deal with discrete-valued data are inherently more appropriate.

One such method is based on the work of Grizzle, Starmer, and Koch (1969) and can be carried out via the computer program GENCAT. This method is not a fully automated procedure, so the variable selection problem is essentially a separate problem from that of model fitting. The combined operation of variable selection and model fitting, however, is analogous to stepwise multiple regression analysis for discrete-valued or categorical data. Because of this categorical orientation, the GENCAT procedure was chosen as the most appropriate for this project.

An automated procedure - AID (Automatic Interaction Detection) was also used as an alternative method of variable selection. This method did not prove to be very satisfactory for the present application, however. Details of comparisons of these two variable selection methods are presented in the section on variable selection.

5.3 Selection of Variable Levels and Time Frames

Table 5.1 shows a list of the variables that were available on the data files for possible inclusion as predictor variables for A/R crashes. As noted in the footnote for this table, the values of each of the driver history variables were accumulated over as many as eight six-month intervals; thus, it was necessary to select the most appropriate time frame for these variables for each high-risk group. It was also necessary to select the levels or value ranges for nearly all the variables to be used in the modeling procedure. The

Table 5.1 Variables examined in variable selection process.

I. Demographic variables

1. Age
2. Sex
3. Race
4. Divorce
5. Prison

II. Accident variables

6. Total crashes
7. Total A/R crashes
8. Total night crashes
9. Time of week
10. Locality
11. Weather
12. Severity
13. Accident type
14. Occupants
15. Type of violation

III. Driver history variables*

16. No. of speeding convictions (or violations)
17. No. of stop convictions (or violations)
18. No. of moving convictions (or violations)
19. No. of reckless convictions (or violations)
20. No. of alcohol convictions (or violations)
21. No of administrative convictions (or violations)
22. No. of accidents at fault
23. No. of suspension & revocation violations
24. No. of equipment violations
25. Total violations
26. Total accidents
27. Total 4-point letters
28. Total 7-point letters
29. Total suspensions
30. Total revocations
31. Total conferences
32. Total hearings
33. Total preliminary hearings
34. Total accidents not at fault
35. Total days of suspension and/or revocation

IV. Alcohol-related arrest variables (Raters variables)

36. No. of violations
37. No. of day violations
38. No. of night violations
39. Blood alcohol concentration
40. No. of crash involved arrests
41. No. of DUI's tried
42. No. of other offenses tried
43. No. of DUI convictions
44. No. of other convictions
45. No. of not guiltys for noted offense
46. No. of prayers for judgment continued
47. No. of nol pros's

*The values of the driver history variables are accumulated over six month intervals for eight such intervals thus there is a choice of the best time frame for each group.

process of selecting the time frames and levels is illustrated below for two of the driver history variables with respect to the general population sample.

Table 5.2 shows five two-way contingency tables of A/R crash involvement versus the number of days under suspension/revocation for different time intervals ranging from the last six months prior to 1975 to the last four years prior to 1975. Several trends can be noted from this table. From the rightmost column it is seen that the overall sample decreases as longer histories are used (since complete records are not available for all drivers). It may also be noted from this column that the overall proportion of drivers having an A/R crash decreases slightly (from .40 to .36) as the longer time frame is used. From the body of the table, it can be seen that, as the length of record increases, the number of drivers having A/R crashes and suspensions/revocations increases (as would be expected), and that these drivers are distributed toward the right hand side of the table (i.e., in those columns corresponding to longer suspensions/revocations). The percentage of drivers having A/R crashes is clearly different for drivers having no days of suspension/revocation as compared with those having one or more days of suspension/revocation, as can be seen by comparing the first column with the other columns. On the other hand, differences in these percentages across the various number of days categories are not so clearly defined. This fact, together with the relatively small numbers in most of the A/R crash cells, indicates that the most appropriate categorization of this variable would be to have two levels corresponding to drivers having no days of suspension/revocation and those having one or more days of suspension/revocation.

Table 5.3 shows the number and percentage in the combined category of those drivers having an A/R crash and some days of suspension/revocations as a function of the length of the record being used. From this table it can be seen

Table 5.2. Contingency table of 1975 alcohol related crashes versus number of days under suspension/revocation.

Number of days under suspension/revocation

	0	1-30	31-60	61-90	91-184	185+	Total
One or More A/R Crashes	796 (.35)	24 (2.55)	22 (2.33)	12 (1.80)	82 (1.32)	0	936 (.40)
No A/R Crashes	221168 (99.65)	914 (97.45)	919 (97.67)	652 (98.20)	6105 (98.68)	1 (100.00)	229759 (99.60)
Last six months prior to 1975							230695

	0	1-30	31-60	61-90	91-184	185+	Total
One or More A/R Crashes	698 (.32)	18 (2.10)	24 (2.22)	6 (.93)	49 (3.15)	86 (1.42)	881 (.39)
No A/R Crashes	210842 (99.68)	837 (97.89)	1053 (97.78)	636 (99.07)	1505 (96.85)	5947 (98.58)	220820 (99.61)
Last year prior to 1975							221701

	0	1-30	31-60	61-90	91-184	185+	Total
One or More A/R Crashes	545 (.28)	24 (2.26)	22 (1.64)	9 (1.24)	50 (3.08)	142 (1.87)	792 (.38)
No A/R Crashes	191719 (99.72)	1037 (97.74)	1317 (98.35)	716 (98.75)	1573 (96.92)	7464 (98.13)	203826 (99.62)
Last two years prior to 1975							204618

	0	1-30	31-60	61-90	91-184	185+	Total
One or More A/R Crashes	441 (.24)	18 (1.56)	22 (1.56)	13 (1.74)	43 (2.70)	183 (2.06)	720 (.37)
No A/R Crashes	176414 (99.76)	1129 (98.44)	1370 (98.44)	738 (98.26)	1550 (97.30)	8736 (97.94)	189937 (99.63)
Last three years prior to 1975							190657

	0	1-30	31-60	61-90	91-184	185+	Total
One or More A/R Crashes	347 (.21)	13 (1.14)	28 (1.92)	10 (1.31)	38 (2.62)	206 (2.11)	642 (.36)
No A/R Crashes	162446 (99.79)	1128 (98.86)	1430 (98.08)	750 (98.68)	1411 (97.37)	9522 (97.88)	176687 (99.63)
Last four years prior to 1975							177329

Column percentages are shown in parentheses.

Table 5.3. Frequency and percentage of suspended/revoked drivers who experience one or more A/R crashes in 1975 by time interval.

	Number of Suspended/Revoked Drivers Who Have A/R Crashes in 1975	Percentage of Suspended/ Revoked Drivers Who Have A/R Crashes in 1975
Last 6 Months	140	1.60
Last Year	183	1.80
Last 2 Years	247	2.00
Last 3 Years	279	2.02
Last 4 Years	295	2.03

that, as the length of record increases, both the number and percentage also increase. Thus, for this variable it is clear that the longest record (four years) is the best to use.

The set of contingency tables of A/R crashes versus number of reckless driving convictions is presented in Table 5.4. Essentially the same comments that were made relative to Table 5.2 apply here as well. In particular, the big difference in the percentage of drivers having A/R crashes between those with no reckless convictions and those with one or more, together with the very small numbers in the cells corresponding to A/R crashes and two or more reckless convictions, indicates that it would again be appropriate to consider only two levels of the variable--no reckless convictions, and one or more reckless convictions.

Table 5.5 gives the number and percentage of drivers having one or more A/R crashes and one or more reckless convictions as a function of record length. Here the choice of optimal record length is not so clear-cut, since, although the number of drivers increased with increasing record length, the percentage having A/R crashes decreased. Whereas in the case of the variable "total days of suspension/revocation" it was possible to choose a record length which simultaneously maximized the cell size and percentage, this is clearly not possible with the variable "number of reckless convictions". Since the models would usually take several variables into account simultaneously, it seemed most important to have as large a sample size as possible as long as the corresponding proportion of drivers having A/R crashes did not become too small. As a rule of thumb, in situations like this second example, the time frame that was chosen maximized the product of the sample size and the percentage.

These two examples illustrate the considerations that went into the selection of time frames and variable levels for each variable with respect to each high-risk group.

Table 5.4. Contingency tables of alcohol related crashes versus number of reckless driving convictions.

	Number of reckless driving convictions						
	0	1	2	3	4	5	Total
One or More A/R Crashes	899 (.39)	37 (2.88)	0	0	0	0	936 (.40)
No A/R Crashes	228493 (99.60)	1245 (97.12)	20 (100.00)	1 (100.00)	0	0	229759 (99.60)
			Last six months prior to 1975				221701

	0	1	2	3	4	5	Total
One or More A/R Crashes	820 (.37)	58 (2.39)	3 (3.57)	0	0	0	881 (.39)
No A/R Crashes	218373 (99.62)	2360 (97.60)	81 (96.42)	6 (100.00)	0	0	220820 (99.61)
			Last year prior to 1975				221701

	0	1	2	3	4	5	Total
One or More A/R Crashes	696 (.34)	85 (1.95)	10 (4.04)	1 (4.54)	0	0	792 (.38)
No A/R Crashes	199311 (99.66)	4257 (98.05)	237 (95.06)	21 (95.45)	0	0	203826 (99.62)
			Last two years prior to 1975				204618

	0	1	2	3	4	5	Total
One or More A/R Crashes	597 (.32)	108 (1.80)	13 (2.80)	2 (3.92)	0	0	720 (.37)
No A/R Crashes	183567 (99.68)	5866 (98.20)	450 (97.20)	49 (96.08)	4 (100.00)	0	189937 (99.63)
			Last three years prior to 1975				190657

	0	1	2	3	4	5	Total
One or More A/R Crashes	500 (.29)	118 (1.67)	21 (3.24)	3 (3.61)	0	0	642 (.36)
No A/R Crashes	169036 (99.71)	6932 (98.33)	626 (96.76)	80 (96.39)	10 (100.00)	3 (100.00)	176687 (99.64)
			Last four years prior to 1975				177329

Table 5.5. Number and percentage of drivers from general population sample having one or more A/R crashes in 1975 and one or more reckless driving convictions in the indicated time interval.

	Number of "Reckless" Drivers With A/R Crashes	Percentage of Total "Reckless" Drivers With A/R Crashes
Last 6 Months	37	2.84
Last Year	61	2.43
Last 2 Years	96	2.08
Last 3 Years	123	1.89
Last 4 Years	142	1.82

5.4 Stepwise Selection of Variables.

After the time frame and levels for each variable had been chosen for every high-risk group, a stepwise procedure was used to select, for each group, a subset of variables to use in a model for predicting that group's A/R crash involvement. At the first step in the procedure, two-way contingency tables similar to those of Tables 5.2 and 5.4 were constructed for each potential independent variable vs A/R crashes. If any variables were significantly related to A/R crashes, as indicated by a x^2 statistic, the variable with the highest value of x^2 divided by degrees of freedom was the variable selected at this first step. Thus, x^2/d.f. served as a measure of the variation in the likelihood of an A/R crash that could be accounted for by the independent variable. The variables selected at the first step for each of the seven groups were as follows:

Group	Variable
General population	Total days under suspension/revocation in last four years
16-20 year old males	Total days under suspension/revocation in last year
21-24 year old males	Total days under suspension/revocation in last four years
DUI	Driver age
Three or more violations	Driver age
Divorce	Number of alcohol-related convictions in last four years
Prison	Number of administrative violation convictions in last four years

The second step in the variable selection procedure was to select another variable for each group which contributed the most toward the prediction

of A/R crashes beyond that contributed by the first variable. To do this, three-way contingency tables were analyzed for each of the remaining variables versus A/R crashes and the variables selected in the first step. Table 5.6 presents an example of such a three-way table which was generated in order to identify a second prediction variable for the 16 to 20-year-old male group. The first variable which had been selected for this group was the total days suspension/revocation variable; the additional variable under consideration was the total number of night crashes. The total variation in the A/R crash rate accounted for by the two variables is again indicated by the x^2/d.f. A test of significance of the second variable is obtained as the sum of the x^2 statistics computed for each of the two-way tables of A/R crashes vs the second variable defined by each level of the first variable. Thus, in Table 5.6 there are two such subtables of A/R crashes vs total night crashes corresponding to the two levels of the days suspension/revocation variable. Their respective x^2's and the sum are shown at the bottom of the table. The second variable then, is selected by identifying that variable which, together with the variable previously selected, accounts for the largest variation in the A/R crash rate, and then checking its statistical significance. If the variable is significant, it becomes the one selected at this step; if not, then the one accounting for the next largest amount of variation is tested, and so on. Of course, if no variable is significant, then none is selected and the procedure is terminated.

The selection of additional variables follows very much the same sort of procedure. One important difference, however, concerns the significance test that is used. After the data have been partitioned by several variables, the numbers of observations in some cells of the resulting multi-way contingency tables may become so small that the x^2 statistics for some of the subtables

Table 5.6. Three-way contingency table of A/R crashes versus total days suspension/revocation and total night crashes - 16-20 year old male group.

	No Days Suspension/Revocation		One or More Days Suspension/Revocation		
	No Night Crashes	One or More Night Crashes	No Night Crashes	One or More Night Crashes	Total
One or More A/R Crashes	855 (1.09)	125 (2.10)	173 (2.67)	73 (4.07)	1226 (1.33)
No A/R Crashes	76890 (98.91)	5805 (97.90)	6300 (97.33)	1717 (95.93)	90712 (98.67)

Overall $\chi^2_{3 \text{ d.f.}} = 249.97$

$\frac{\chi^2}{\text{d.f.}} = 83.32$

No days suspension/revocation $\chi^2_{1 \text{ d.f.}} = 48.38$

One or more days suspension/revocation $\chi^2_{1 \text{ d.f.}} = 9.59$

Chi-square significance test = $\chi^2_{2 \text{ d.f.}} = 57.97$

may be invalid. Moreover, it is usually important that the relationship between the variable being considered and A/R crashes be consistent across the various subtables. For example, if the variable being considered indicates a "good" or "bad" driving record, and if overall drivers with "good" records have lower A/R crash rates, then it is important that this also be true within the levels of the other variables already selected. This kind of consistency may not be required for some variables such as driver age and sex.

A statistic which is valid for subtables with small cell sizes and which emphasizes consistency is the modified Mantel-Haenszel statistic. This was used as the test statistic in the variable selection procedure after the second or third step depending on the overall population size. A general discussion of Mantel-Haenszel procedures can be found in Fleiss (1973) and its use in variable selection is discussed in Clarke & Koch (1974). An illustration of variable consistency and inconsistency and the Mantel-Haenszel statistic is given in Tables 5.7 and 5.8, which were generated relative to determining a fourth predictor variable for the "Three or more violations group." These tables give the five-way contingency tables of A/R crashes vs respectively, stop sign violations and night crashes within the levels of the three variables already selected at this point. The three previously selected variables are:

1. Driver age - (under 21), (21 and older),
2. Total days sus/rev. - (0 days), (1 or more days),
3. Total A/R crashes (73-74) - (no crashes), (1 or more crashes).

The eight 2 x 2 subtables of Tables 5.7 and 5.8 correspond to all combinations of the levels of these three variables. Thus, subtable 1 corresponds to young drivers having no days suspension/revocation and no A/R crash in 73-74, subtable 2 to young drivers with no days suspension/revocation, but with one or more A/R crashes in 73-74, subtable 3 to young drivers having some suspension/revocation

but no A/R crashes, and so on. In Table 5.7 it can be seen that in three of the eight subtables, including the one with the largest frequencies, the A/R crash rate (percentage) is higher for drivers with no stop sign violations than it is for those with one or more such violations, and, hence, the relationship between A/R crashes and stop violations is not very consistent over the levels of the other important variables. This fact is reflected in the Mantel-Haenszel statistic, which is quite small and very nonsignificant.

On the other hand, for each of the subtables of Table 5.8, the A/R crash rate is always higher for drivers having one or more night crashes than for those having none. While the overall variation in A/R crash rates as indicated by the x^2/d.f. statistic is only slightly higher for this table, the Mantel-Haenszel statistic is highly significant for Table 5.8 indicating a strong consistent relationship between A/R crashes and night crashes even after the other three variables have been taken into account.

The variable selection procedure was terminated either when no more significant variables remained, or when the data had been partitioned to the extent that the high-risk subgroups contained so few individuals that further subdivision was not feasible. In the latter case, it may be that additional variables remain which are significantly related to the A/R crash variable beyond those included in the model. The effect of these additional variables is, however, only to further partition the lower risk subgroups into other subgroups with different but still relatively low A/R crash rates. Thus, these variables do not contribute to the prediction of the higher risk subgroups, and, in fact, their inclusion would result in many of the higher risk subgroups having very few or no drivers in the A/R crash cells.

The results of the variable selection procedure applied to each of the high-risk groups are given in Table 5.9. This table shows the set of variables

Table 5.7. Five-way contingency table for stop sign violations.

	1 (<21, No S/R, No 73-74 A/R Crash)		2 (<21, No S/R, ≥1 73-74 A/R Crash)	
	No Stop Violations	One or More Stop Violations	No Stop Violations	One or More Stop Violations
One or More A/R Crashes	320 (3.78)	39 (3.33)	20 (5.55)	4 (7.85)
No A/R Crashes	8132 (96.22)	1132 (96.67)	340 (94.45)	47 (92.15)

	3 (<21, Some S/R, No 73-74 A/R Crash)		4 (<21, Some S/R, ≥1 73-74 A/R Crash)	
	No Stop Violations	One or More Stop Violations	No Stop Violations	One or More Stop Violations
One or More A/R Crashes	167 (4.40)	35 (4.42)	28 (7.60)	4 (5.48)
No A/R Crashes	3623 (95.60)	758 (95.58)	340 (92.40)	69 (94.52)

	5 (≥21, No S/R, No 73-74 A/R Crash)		6 (≥21, No S/R, ≥1 73-74 A/R Crash)	
	No Stop Violations	One or More Stop Violations	No Stop Violations	One or More Stop Violations
One or More A/R Crashes	1101 (1.68)	100 (1.55)	154 (5.28)	20 (5.84)
No A/R Crashes	64070 (98.32)	6343 (98.45)	2762 (94.72)	322 (94.16)

	7 (≥21, Some S/R, No 73-74 A/R Crash)		8 (≥21, Some S/R, ≥1 73-74 A/R Crash)		
	No Stop Violations	One or More Stop Violations	No Stop Violations	One or More Stop Violations	Total
One or More A/R Crashes	670 (2.24)	53 (2.71)	110 (2.90)	12 (4.02)	2837 (2.25)
No A/R Crashes	29200 (97.76)	1902 (97.29)	3683 (97.10)	286 (95.98)	123009 (97.75)

Overall $\frac{\text{Chi-square}}{\text{15 degrees of freedom}} = 35.48$

Mantel-Haenszel $\chi^2_{1\text{d.f.}} = .0009$

Table 5.8. Five-way contingency table for night crashes.

1 (<21, No S/R, No 73-74 A/R Crash)

	No Night Crashes	One or More Night Crashes
One or More A/R Crashes	310 (3.70)	49 (3.96)
No A/R Crashes	8073 (96.30)	1191 (96.04)

2 (<21, No S/R, ≥1 73-74 A/R Crash)

	No Night Crashes	One or More Night Crashes
One or More A/R Crashes	2 (3.28)	22 (6.28)
No A/R Crashes	59 (96.72)	328 (93.72)

3 (<21, Some S/R, No 73-74 A/R Crash)

	No Night Crashes	One or More Night Crashes
One or More A/R Crashes	154 (4.06)	48 (6.12)
No A/R Crashes	3644 (95.94)	737 (93.88)

4 (<21, Some S/R, ≥1 73-74 A/R Crash)

	No Night Crashes	One or More Night Crashes
One or More A/R Crashes	2 (2.38)	30 (8.40)
No A/R Crashes	82 (96.72)	327 (91.60)

5 (≥21, No S/R, No 73-74 A/R Crash)

	No Night Crashes	One or More Night Crashes
One or More A/R Crashes	1057 (1.60)	144 (2.54)
No A/R Crashes	64900 (98.40)	5513 (97.46)

6 (≥21, No S/R, ≥1 73-74 A/R Crash)

	No Night Crashes	One or More Night Crashes
One or More A/R Crashes	49 (4.99)	125 (5.48)
No A/R Crashes	932 (95.01)	2152 (94.52)

7 (≥21, Some S/R, No 73-74 A/R Crash)

	No Night Crashes	One or More Night Crashes
One or More A/R Crashes	664 (2.22)	59 (2.92)
No A/R Crashes	29147 (97.78)	1955 (97.08)

8 (≥21, Some S/R, ≥1 73-74 A/R Crash)

	No Night Crashes	One or More Night Crashes	Total
One or More A/R Crashes	40 (2.68)	82 (3.16)	2837 (2.25)
No A/R Crashes	1448 (97.32)	2521 (96.84)	123009 (97.75)

Overall $\frac{\text{Chi-square}}{\text{15 degrees of freedom}} = 38.51$

Mantel-Haenszel $\chi^2_{1d.f.} = 32.08$

Table 5.9. Variables selected for A/R crash prediction models.

Group	Variable	Levels
General population (I)	1. Total days S/R (4 yrs.)	none, one or more
	2. Accident violations (4 yrs.)	none, one or more
	3. Sex	M, F
	4. Reckless violations (4 yrs.)	none, one or more
General population (II)	1, 2, 3, as in (I)	
	4. Age	under 25, 25 and over
16-20 yr. old males	1. Total days S/R (1 yr.)	none, one or more
	2. Total violations (1 yr.)	none, one or more
	3. Night crashes (73-74)	none, one or more
	4. Night violation arrests (1 yr.)	none, one or more
21-24 yr. old males	1. Total days S/R (4 yrs.)	none, one or more
	2. Reckless violations (4 yrs.)	none, one or more
	3. Alcohol violations (4 yrs.)	none, one or more
	4. A/R crashes (73-74)	none, one or more
DUI (I)	1. Driver age	20 or under, 21-25, over 25
	2. Speeding violations (1 yr.)	none, one or more
	3. Total days S/R (3 yrs.)	less than 185, 185 or more
	4. Reckless violations (1 yr.)	none, one or more
DUI (II)	1. Driver age	25 or less, over 25
	2, 3, 4 as above in (I)	
DUI (III)	1, 2, 3 as in (II)	
	4. Accidents not at fault (3 yrs.)	none, one or more
Three or more violations	1. Age	under 21, 21 and over
	2. Total days S/R (1 yr.)	none, one or more
	3. A/R crashes (73, 74)	none, one or more
	4. Night crashes (73, 74)	none, one or more
	5. Sex	M, F
Divorce	1. Alcohol violations (4 yrs.)	none, one or more
	2. Reckless violations (3 yrs.)	none, one or more
Prison	1. Administrative viol. (2 yrs.)	none, one or more
	2. Age of driver	30 or under, over 30

selected for each high-risk group, with the number of the variable indicating the order in which the variables were selected. Accompanying each of the driver history variables is the time span over which the number of events was accumulated (e.g., the number of years prior to 1975). The variables labelled as violations (i.e., reckless violations, alcohol violations, etc.) refer to counts of convictions for these violations. In contrast, the variable "night violation arrests" appearing in the set of variables for the 16-20 year old male group refers to arrests which may or may not have resulted in convictions.

For some high-risk groups (e.g. DUI), it will be noted that more than one set of variables is listed in Table 5.9. This resulted from the fact that at some stage in the variable selection procedure, a clear choice between two variables could not be made. When this happened, both variables were carried forward through the remaining steps of the selection procedure. Models were later developed for each set of variables and a choice was made between the various sets of variables on the basis of the model coefficients, predicted values, goodness of fit tests, and concurrent validity tests.

As an alternative method of variable selection, the Automatic Interaction Detection (AID) procedure was applied to the data depicted in Table 5.1 for the 16 to 20-year-old male group. The subgroups identified by AID were defined in terms of the following four variables:

1.	Total days S/R (1 yr.)	none, one or more
2.	Total violations (1 yr.)	none, one, two or more
3.	Night violation arrests (1 yr.)	none, one or more
4.	Speeding violations (1 yr.)	none, one or more

Thus, the first three variables selected by AID are included in the set shown in Table 5.9. As the fourth variable, the AID procedure selected speeding violations, whereas the GENCAT procedure selected night crashes. The AID

procedure is a fully automated algorithm which does not take into account any sort of consistency restrictions concerning the relationships between the variables. The speeding violations variable turned out to be highly inconsistent (in the sense described in the preceding pages) within the levels of the other variables. In fact, speeding had an overall negative effect in that most of the time drivers with no speeding violations had higher crash rates than did those with one or more. The highest crash rate was for drivers who had had two or more violations and one or more night violation arrests, but no days S/R, and no speeding violations. This rate was nearly twice as high as that for drivers also having two or more violations, one or more night violation arrests, one or more days S/R, and one or more speeding violations. In view of these results and the fact that AID was very costly to use with large data files, it was decided, then, that further use of the AID program was unlikely to provide usable information, and therefore, would not be cost effective.

An examination of Table 5.9 reveals that many of the same variables appear in the selected sets for several different groups. Age and sex are used to define the two young male groups. Age also appears as an important variable for each of the other groups except the divorce group (which contained few very young drivers) while sex appears in two other groups. The total days of suspension/revocation appears to be a very powerful predictor variable and was selected for all groups except the rather small Prison and Divorce groups. Reckless violations also seem quite important and were selected for four of the seven groups. Other variables that were selected for more than one group include alcohol violations, night crashes, and A/R crashes.

5.5 Model Fitting

After predictor variables were determined for each of the high-risk groups, categorical data models could be developed to predict A/R crash rates in terms of these variables. As mentioned earlier, the stepwise variable selection followed by the fitting of linear categorical models is exactly analogous to forward stepwise regression analysis in the continuous variable case. The final crosstabulations from the variable selection phase provided the definitions of a set of categories or subpopulations together with frequencies and proportions of the occurrence of A/R crashes for each subpopulation. For example, four variables, each of which had two levels, were selected for the general population group. The combinations of these levels generated sixteen distinct subpopulations. Table 5.10 shows these subpopulations together with their respective A/R crash frequencies, proportions, and the standard errors of the proportions. Thus, the first subpopulation corresponds to males with no days suspension/revocation, no accident violations, and no reckless violations. The proportion of the 77,701 drivers in this subpopulation who had A/R crashes in 1975 was .00281.

Linear categorical models were then fit to the resulting column of observed proportions for each set of variables (more than one set of variables having been chosen for some high-risk groups). These models are of the general form

$$E(P) = XB.$$

where P is the vector of subpopulation A/R crash proportions, X is a design matrix whose columns represent effects due to the variables and their interactions, and B is a vector of model coefficients to be estimated. A discussion of these models can be found in Grizzle, Starmer, and Koch (1969).

As a starting point for the development of such a model, a basic design matrix may be specified in a variety of ways. A basic form which usually provided a good starting point for models with sixteen subpopulations is shown

Table 5.10. Subpopulations and AR crash frequencies - general population group.

	Subpopulations				Frequencies		P	
	Total Days S/R	Accident Violations	Driver Sex	Reckless Violations	No A/R Crashes	One or More A/R Crashes	Observed Proportions	Standard Errors of Proportions
1.	N	N	M	N	77483	218	.00281	.00019
2.	N	N	M	S	2986	33	.01093	.00189
3.	N	N	F	N	72794	36	.00049	.00008
4.	N	N	F	S	538	2	.00370	.00261
5.	N	S	M	N	4509	40	.00879	.00138
6.	N	S	M	S	1030	16	.01530	.00379
7.	N	S	F	N	2821	2	.00071	.00050
8.	N	S	F	S	285	0	0	.00247*
9.	S	N	M	N	8762	143	.01606	.00133
10.	S	N	M	S	1603	53	.03201	.00433
11.	S	N	F	N	913	1	.00109	.00109
12.	S	N	F	S	64	1	.01539	.01527
13.	S	S	M	N	1595	58	.03509	.00453
14.	S	S	M	S	1093	36	.03189	.00523
15.	S	S	F	N	159	2	.01242	.00873
16.	S	S	F	S	52	1	.01887	.01869

*Standard error computed with 0.0 frequency replaced with 0.5.

N ≡ none, S ≡ one or more, M ≡ male, F ≡ female

in Figure 5.2. This matrix is in block diagonal or modular form. The first module (first four rows and first three columns) corresponds (from Table 5.10), to drivers with no days suspension/revocation and no accident violations, the second to drivers with no days suspension/revocation but some accident violations, the third to those with some days suspension/revocation but no accident violations, and the fourth to drivers with both days suspension/revocation and accident violations. The first column (column of ones) within each module represents a baseline effect for that module. The second column represents a sex effect, and the third, an effect due to the reckless violations variable. A vector of twelve regression coefficients (the number of columns in the design matrix) is estimated for the model by the method of weighted least squares. If the goodness of fit statistic is not significant, indicating that the model provides an adequate representation of the data, then a series of hypotheses on the values of the parameters can be tested. In particular, it is of interest to test hypotheses that coefficients of the same variables (e.g., sex effects) in different modules are equal. When the test statistics for these tests are not significant, certain columns of the basic design matrix can be combined, and at times, others can be deleted. The objective of these hypothesis tests is to obtain a reduced (and, hence, simpler) design matrix which has fewer columns but still provides an adequate fit to the observed proportions.

The observed proportion for a given subpopulation is determined from the A/R crash frequencies for that subpopulation only, as are the estimated standard deviations or standard errors. The model provides estimated or predicted proportions, however, that are determined from the frequencies from all of the subpopulations. Thus, in effect, the model "smooths" the raw proportions to yield the predicted ones. The standard errors of the predicted proportions are, hence, usually much smaller than those of the raw proportions.

Figure 5.2. Basic design matrix.

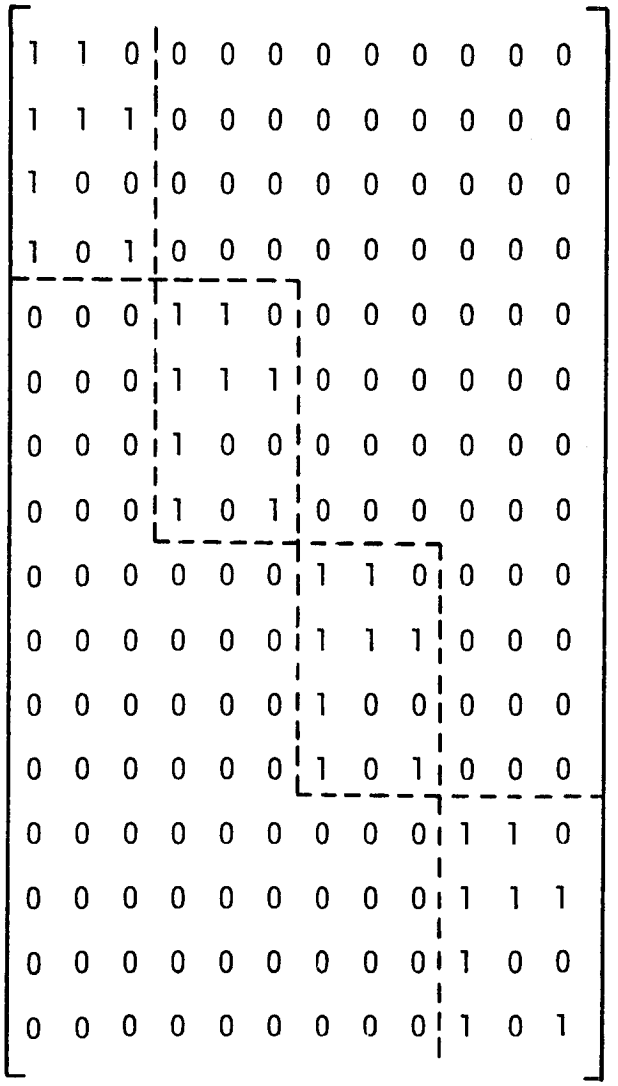


Table 5.11 contains the same information as Table 5.10, but with the addition of two more columns containing the predicted proportions and their standard errors. The predicted proportions are also shown in Figure 5.3 to illustrate the wide range of variation in these values across the subpopulations. Similar figures for all high-risk groups appear in Appendix B.

Figure 5.4 shows the reduced design matrix and the vector of estimated model coefficients which together generate the predicted values of Table 5.11. The predicted values are obtained by the matrix multiplication

$$P = XB \quad ,$$

where P is the vector of predicted A/R crash proportions, X is the reduced design matrix, and B is the vector of model coefficients. For example, the first predicted value is given by

$$P_1 = .00050 + .00234 = .00284 \quad ,$$

the second by

$$P_2 = .00050 + .00234 + .00489 = .00773 \quad ,$$

and so forth.

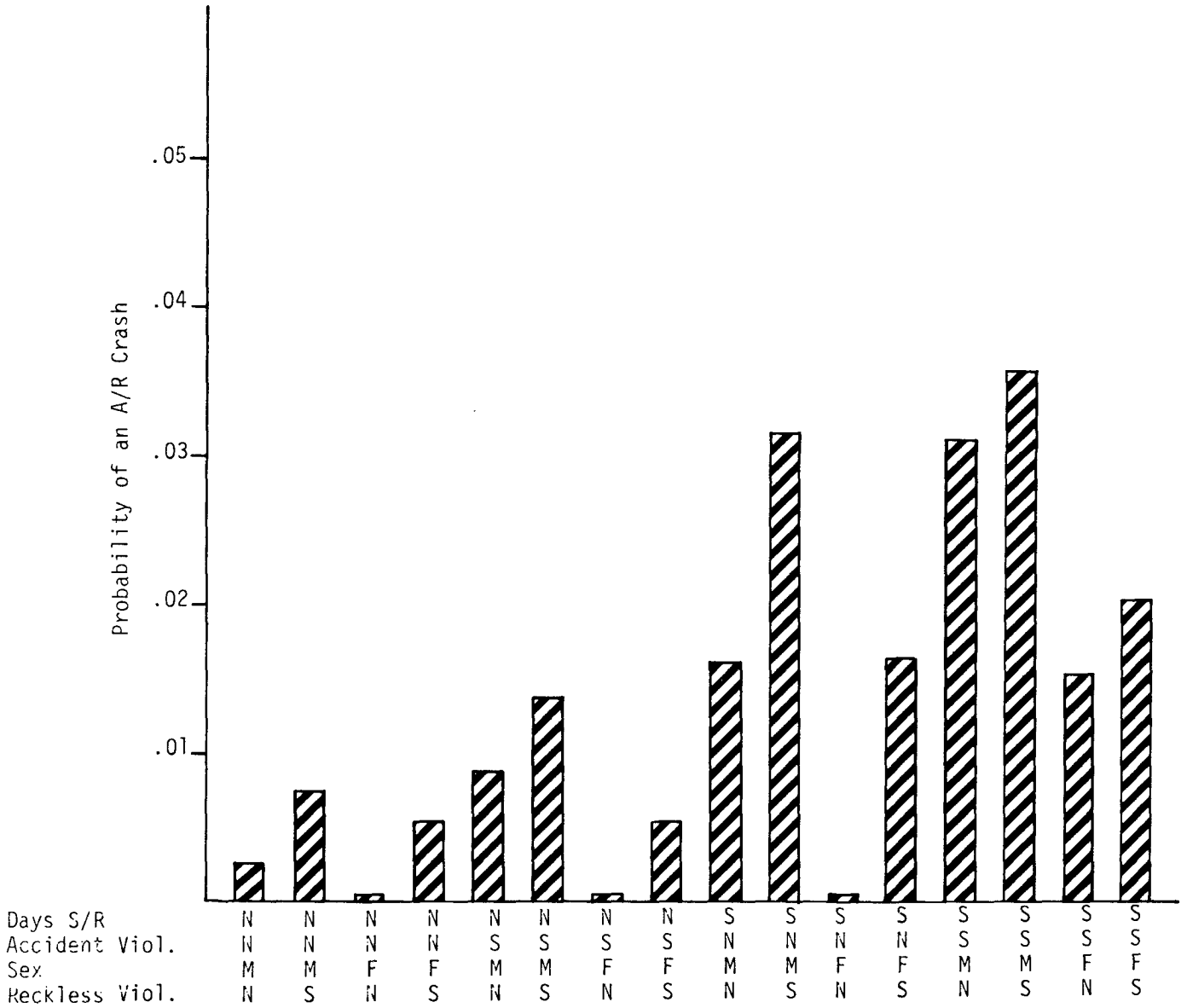
The predicted values shown in Table 5.11 can be seen to be quite close to the raw proportions for most of the subpopulations, especially for those with the larger frequencies (this, of course, is a result of the weighted least squares procedure which gives more weight to those subpopulations with smaller variances or larger frequencies). The standard errors of the predicted proportions in the last column of Table 5.11 are considerably smaller than those for the raw proportions for most subpopulations. Thus, the predicted proportions give more precise estimates of the effects of the variables included in the model than do the raw proportions. This is especially true in the case of subpopulations with very small frequencies.

Table 5.11 Model 1 - General Population Group.

	Days S/R	Acc. Viol.	Sex	Reckless	Frequencies 0 A/R	Frequencies 1+ A/R	Observed Proportions	Standard Error	Predicted Proportions	Standard Error
1.	N	N	M	N	77483	218	.00281	.00019	.00284	.00019
2.	N	N	M	S	2986	33	.01093	.00189	.00773	.00122
3.	N	N	F	N	72794	36	.00049	.00008	.00050	.00008
4.	N	N	F	S	538	2	.00370	.00261	.00539	.00122
5.	N	S	M	N	4509	40	.00879	.00138	.00898	.00131
6.	N	S	M	S	1030	16	.01530	.00379	.01387	.00169
7.	N	S	F	N	2821	2	.00071	.00050	.00050	.00008
8.	N	S	F	S	285	0	0	.00247*	.00539	.00122
9.	S	N	M	N	8762	143	.01606	.00133	.01614	.00131
10.	S	N	M	S	1603	53	.03201	.00433	.03193	.00416
11.	S	N	F	N	913	1	.00109	.00109	.00050	.00008
12.	S	N	F	S	64	1	.01539	.01527	.01629	.00434
13.	S	S	M	N	1595	58	.03509	.00453	.03111	.00318
14.	S	S	M	S	1093	36	.03189	.00523	.03600	.00323
15.	S	S	F	N	159	2	.01242	.00873	.01546	.00336
16.	S	S	F	S	52	1	.01887	.01869	.02035	.00341

*Standard error computed with 0 frequency replaced with 0.5

Figure 5.3. General population model - predicted probabilities of A/R crash involvement.



The columns of the model design matrix shown in Figure 5.4 can be interpreted in terms of the variables used in defining the subpopulations. The first column is made up of the baselines from the first three modules, and hence, is itself a baseline for all driver subpopulations corresponding to drivers having no days of suspension/revocation and no accident violations, or one or the other but not both. The second column is a baseline for subpopulations corresponding to drivers having both days of suspension/revocation and accident violations. The term "baseline" here means that the coefficient of the column gives the minimum probability of an A/R crash for all subpopulations for which a "one" appears in the column. For subpopulations of male drivers and/or drivers with reckless violations, additional probability increments are added to the baseline values. The coefficents of the two baselines described above are .0050 and .01546, respectively. Thus, drivers with both days of suspension/revocation and accident violations have much higher baseline probabilities than those with only one or neither.

The next three columns of the design matrix indicate sex effects in the first, second, and combined third and fourth modules respectively. These columns, together with their coefficients show males always to have higher probabilities of A/R crashes than females, and that the difference between males and females increases with worsening driver record. Finally, the last two columns represent reckless violation effects in the combined first, second, and fourth, and third modules, respectively.

The design matrix shown in Figure 5.4 could be reduced further. Since the second, fifth, and last coefficients are very nearly the same value, it would be possible to combine these columns into a single column which would generate essentially the same predicted probabilities. This has not been done, however,

Figure 5.4 Design matrix and model coefficients - general population model.

1	1	0	1	0	0	0	0
2	1	0	1	0	0	1	0
3	1	0	0	0	0	0	0
4	1	0	0	0	0	1	0
5	1	0	0	1	0	0	0
6	1	0	0	1	0	1	0
7	1	0	0	0	0	0	0
8	1	0	0	0	0	1	0
9	1	0	0	0	1	0	0
10	1	0	0	0	1	0	1
11	1	0	0	0	0	0	0
12	1	0	0	0	0	0	1
13	0	1	0	0	1	0	0
14	0	1	0	0	1	1	0
15	0	1	0	0	0	0	0
16	0	1	0	0	0	1	0

$$B = \begin{bmatrix} .00050 \\ .01546 \\ .00234 \\ .00849 \\ .01565 \\ .00489 \\ .01579 \end{bmatrix}$$

χ^2 - due to model = 469.78 d.f. = 6

χ^2 - due to error = 7.63 d.f. = 9 ($p > .50$)

$R^2 = .984$

Ratio of largest predicted value to smallest = 72.0

since these three columns represent different effects and the combined effect would have a less straightforward interpretation.

This same general approach to model development was followed with respect to the models for the other high-risk groups. In some cases, when the coefficients of different effects were virtually identical, the corresponding columns of the design matrix were combined if such a reduction would not unduly complicate the model.

Below Figure 5.4 are some summary statistics relating to this model. Chi-square statistics for the overall model and the error term are shown together with their respective degrees of freedom. The error term is seen to be very nonsignificant ($p > .50$). An R^2 statistic is obtained as (the x^2 due to model)/(total x^2) and is the proportion of the total variation in A/R crash rates across the sixteen subpopulations that is accounted for by the model. Also shown is the ratio of the largest predicted value to the smallest, which is a measure of the range of predicted values. Thus, while none of the A/R crash rates is large in the absolute sense, the rate of the "worst" group is seventy-two times as high as that of the "best" groups. Design matrices for all groups appear in Appendix C.

5.6 The Models

In general, all of the models provided good fits to the data. This can be seen by examining the predicted and actual proportions of drivers having A/R crashes in Tables 5.12-5.17. Also shown on these tables are statistics which indicate the goodness-of-fit of the models. These include the x^2 due to error statistics, all of which are highly nonsignificant ($p \geq .50$ for all groups except the divorce and prison groups where the error terms have only one degree of freedom). The R^2 statistic, another measure of goodness of fit, is well above .90 for all groups except the DUI group and the prison group. With

Table 5.12 Model 2 - Males, 16-20.

	Total Days S/R	Total Violations	Night Crashes	Night Viol. Arrests	Frequencies: No A/R Crashes	Frequencies: 1+ A/R Crashes	Observed Proportions	Standard Errors	Predicted Proportions	Standard Errors
1.	N	N	N	N	61021	579	.00940	.00039	.00933	.00039
2.	N	N	N	S	123	0	0	.00571*	.00933	.00039
3.	N	N	S	N	3467	58	.01645	.00214	.01788	.00172
4.	N	N	S	S	28	1	.03448	.03388	.01788	.00172
5.	N	S	N	N	15444	258	.01643	.00101	.01664	.00098
6.	N	S	N	S	302	18	.05625	.01288	.03956	.00635
7.	N	S	S	N	2203	59	.02608	.00335	.02519	.00186
8.	N	S	S	S	107	7	.06140	.02248	.04810	.00647
9.	S	N	N	N	1787	53	.02880	.00390	.02533	.00187
10.	S	N	N	S	11	0	0	.06014*	.04824	.00622
11.	S	N	S	N	446	17	.03672	.00874	.03387	.00238
12.	S	N	S	S	11	0	0	.06014*	.05679	.00632
13.	S	S	N	N	3973	94	.02311	.00236	.02533	.00187
14.	S	S	N	S	529	26	.04685	.00897	.04824	.00622
15.	S	S	S	N	1078	49	.04348	.00608	.03387	.00238
16.	S	S	S	S	182	7	.03704	.01374	.05679	.00632

*Standard errors computed with 0 frequency replaced with 0.5.

χ^2 due to model = 185.40 d.f. = 4

χ^2 due to error = 10.14 d.f. = 11 $p > .50$

$R^2 = .948$

Ratio of largest predicted value to smallest = 6.09

Table 5.13 Model 3 - Males, 21-24.

	Total Days S/R	Reckless Viol.	Alcohol Viol.	A/R Crashes	Frequencies: No A/R Crashes	Frequencies: 1+ A/R Crashes	Observed Proportions	Standard Errors	Predicted Proportions	Standard Errors
1.	N	N	N	N	37415	516	.01360	.00060	.00698	.00031
2.	N	N	N	S	715	16	.02189	.00541	.02051	.00207
3.	N	N	S	N	252	5	.01946	.00862	.02051	.00207
4.	N	N	S	S	28	1	.03448	.03388	.03404	.00412
5.	N	S	N	N	7746	134	.01701	.00146	.01620	.00092
6.	N	S	N	S	399	23	.05450	.01105	.04804	.00674
7.	N	S	S	N	62	1	.01587	.01575	.02973	.00204
8.	N	S	S	S	19	1	.05000	.04873	.06157	.00690
9.	S	N	N	N	9764	154	.01553	.00124	.01620	.00092
10.	S	N	N	S	215	10	.04444	.01374	.02973	.00204
11.	S	N	S	N	3156	100	.03071	.00302	.02973	.00204
12.	S	N	S	S	354	12	.03279	.00931	.04326	.00399
13.	S	S	N	N	4966	112	.02206	.00206	.02240	.00192
14.	S	S	N	S	421	24	.05393	.01071	.05424	.00667
15.	S	S	S	N	1313	54	.03950	.00527	.03593	.00260
16.	S	S	S	S	299	19	.05975	.01329	.06777	.00681

χ^2 due to model = 345.38 d.f. = 4

χ^2 due to error = 5.22 d.f. = 11 $p > .90$

$R^2 = .985$

Ratio of largest predicted value to smallest = 9.71

Table 5.14. Model 4 - DUI Group.

	Driver Age	Speeding Viol.	Days S/R	Reckless Viol.	Frequencies 0A/R crash	Frequencies 1+A/R crash	Observed Proportions	Standard Error	Predicted Proportions	Standard Error
1.	1	N	N	N	243	5	.02016	.00893	.02477	.00256
2.	1	N	N	S	35	1	.02778	.02739	.03087	.00460
3.	1	N	S	N	437	27	.05819	.01087	.05830	.00981
4.	1	N	S	S	33	3	.08333	.04606	.06440	.01057
5.	1	S	N	N	100	2	.01961	.01373	.03738	.00381
6.	1	S	N	S	19	0	0	.03579*	.04348	.00504
7.	1	S	S	N	67	4	.05634	.02736	.07091	.01024
8.	1	S	S	S	17	3	.15000	.07984	.07701	.01081
9.	2	N	N	N	990	27	.02655	.00504	.02477	.00256
10.	2	N	N	S	106	4	.03636	.01785	.03087	.00460
11.	2	N	S	N	3010	96	.03091	.00311	.02919	.00242
12.	2	N	S	S	160	11	.06433	.01876	.03530	.00460
13.	2	S	N	N	309	13	.04037	.01097	.03738	.00381
14.	2	S	N	S	56	0	0	.01246*	.04348	.00504
15.	2	S	S	N	313	14	.04281	.01119	.04180	.00385
16.	2	S	S	S	44	3	.06383	.03566	.04790	.00514
17.	3	N	N	N	7200	111	.01518	.00143	.01507	.00134
18.	3	N	N	S	423	12	.02759	.00785	.02118	.00421
19.	3	N	S	N	21045	410	.01911	.00093	.01950	.00090
20.	3	N	S	S	578	15	.02530	.00645	.02560	.00418
21.	3	S	N	N	809	27	.03230	.00611	.02768	.00350
22.	3	S	N	S	86	3	.00371	.01913	.03378	.00495
23.	3	S	S	N	1028	40	.03745	.00581	.03211	.00351
24.	3	S	S	S	113	5	.04237	.01854	.03821	.00502

*Standard errors computed with 0 frequency replaced with 0.5

χ^2 due to model = 61.28 Ratio of largest predicted value to smallest = 5.11

χ^2 due to error = 16.73 d.f. = 18 $p > .50$ $R^2 = .786$

Table 5.15. Model 5 - Three or more violations.

	Age	Sex	Days S/R	A/R Crashes	Night Crashes	Frequencies 0 A/R	Frequencies 1+ A/R	Observed Proportions	Standard Errors	Predicted Proportions	Standard Errors
1.	Y	M	N	N	N	7294	296	.03900	.00222	.03946	.00175
2.	Y	M	N	N	S	1090	47	.04134	.00590	.04620	.00226
3.	Y	M	N	S	N	54	2	.03571	.02480	.04288	.00316
4.	Y	M	N	S	S	320	21	.06158	.01302	.04962	.00305
5.	Y	M	S	N	N	3430	151	.04217	.00336	.03946	.00175
6.	Y	M	S	N	S	700	48	.06417	.00896	.06780	.00768
7.	Y	M	S	S	N	81	2	.02410	.01683	.04288	.00316
8.	Y	M	S	S	S	318	28	.08092	.01466	.07122	.00789
9.	Y	F	N	N	N	779	14	.01765	.00468	.01739	.00052
10.	Y	F	N	N	S	101	2	.01942	.01360	.02413	.00158
11.	Y	F	N	S	N	5	0	0	.12258*	.02082	.00274
12.	Y	F	N	S	S	8	1	.11111	.10476	.02756	.00264
13.	Y	F	S	N	N	214	3	.01383	.00793	.01739	.00052
14.	Y	F	S	N	S	37	0	0	.01873*	.02413	.00158
15.	Y	F	S	S	N	1	0	0	.38490*	.02082	.00274
16.	Y	F	S	S	S	9	2	.18182	.11629	.02756	.00264
17.	O	M	N	N	N	57509	1011	.01728	.00054	.01739	.00052
18.	O	M	N	N	S	4910	139	.02753	.00230	.02413	.00158
19.	O	M	N	S	N	882	50	.05365	.00738	.05133	.00426
20.	O	M	N	S	S	2051	124	.05701	.00497	.05807	.00415
21.	O	M	S	N	N	27881	644	.02258	.00088	.02255	.00085
22.	O	M	S	N	S	1862	55	.02869	.00381	.02929	.00172
23.	O	M	S	S	N	1384	39	.02741	.00433	.02598	.00261
24.	O	M	S	S	S	2426	80	.03192	.00351	.03272	.00252
25.	O	F	N	N	N	7391	46	.00619	.00091	.00589	.00088
26.	O	F	N	N	S	603	5	.00822	.01306	.01264	.00172
27.	O	F	N	S	N	50	3	.05660	.03174	.00932	.00282
28.	O	F	N	S	S	101	1	.00980	.00976	.01606	.00272
29.	O	F	S	N	N	1266	20	.01555	.00345	.01739	.00052
30.	O	F	S	N	S	93	4	.04124	.02019	.02413	.00158
31.	O	F	S	S	N	64	1	.01539	.01527	.02082	.00274
32.	O	F	S	S	S	95	2	.02062	.01443	.02756	.00264

*Standard error computed with zero frequencies replaced by 0.5

χ^2 due to model = 539.59, d.f. = 7 Ratio of largest predicted value to smallest = 12.09

χ^2 due to error = 16.238, d.f. = 24 $p > .75$ $R^2 = .971$

Table 5.16 Model 6 - Divorce group.

	Alcohol Violations	Reckless Viol.	Frequencies 0 A/R	Frequencies 1+ A/R	Observed Proportions	Standard Error	Predicted Proportions	Standard Error
1.	N	N	7298	42	.00572	.00088	.00570	.00088
2.	N	S	493	10	.01988	.00622	.02118	.00600
3.	S	N	625	22	.03400	.00713	.03571	.00679
4.	S	S	126	9	.06667	.02147	.05119	.00869

χ^2 due to model = 27.53 d.f. = 2

χ^2 due to error = 0.62 d.f. = 1 $p > .25$

$R^2 = .978$

Ratio of largest predicted value to smallest = 8.98

Table 5.17 Model 7 - Prison group.

	Admin. Viol.	Age	Frequencies 0 A/R	Frequencies 1+A/R	Observed Proportions	Standard Error	Predicted Proportions	Standard Error
1.	N	≤30	703	22	.0303	.00637	.0315	.00630
2.	N	>30	1089	21	.0189	.00409	.0184	.00407
3.	S	≤30	75	9	.1071	.03374	.0734	.02028
4.	S	>30	67	3	.0428	.02419	.0602	.01983

χ^2 due to model = 7.62 d.f. = 2

χ^2 due to error = 1.57 d.f. = 1 p = .21

R^2 = .829

Ratio of largest predicted value to smallest = 3.99

these groups, it can be seen that the R^2 is relatively small primarily because the x^2 due to model is much smaller for these models than for any of the other groups. It also may be noted that for these groups the ratios of the largest to smallest predicted values are relatively small. Both of these results stem from the fact that there is relatively little overall variation in A/R crash rates across the subpopulations of these groups. In particular, none of the subpopulations has the very low A/R crash rates that appear in most of the other models. Thus, for example, having a previous DUI arrest seems to guarantee a fairly high A/R crash rate (2.2 percent), and while the other variables have significant effects beyond this, the overall variation in A/R crash rates is relatively low.

The effects of the variables in the models are, in general, very consistent in the sense discussed earlier. One apparent exception to this occurs in the three or more violations group. Here the predicted A/R crash rate for the subgroups of older males with one or more previous A/R crashes, no night crashes, and no days under suspension/revocation is .05133. The corresponding probability for the same no-suspension subgroup with one or more night crashes is .05807. On the other hand, for drivers with the same characteristics but who have a "worse" driver record in that they have one or more days of suspension/revocation, the corresponding rates are only .02598 and .03272. The way that this happens can be seen from an examination of the design matrix in Figure 5.5. The fifth and sixth columns of the design matrix represent previous A/R crash effects. Column five is nearly a main effect in that it indicates the presence or absence of previous A/R crashes for all subpopulations except those in the module defined by older male drivers with no days suspension/revocation. Column six represents the same effect for this module only. The corresponding model coefficients are .00342 and .03393, respectively, so that the presence of previous A/R crashes among the older males with no days of suspension/revocation

Figure 5.5. Design matrix and model coefficients - three or more violations model.

1	1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0
3	1	0	0	0	1	0	0	0
4	1	0	0	0	1	0	1	0
5	1	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	1
7	1	0	0	0	1	0	0	0
8	1	0	0	0	1	0	0	1
9	0	1	0	0	0	0	0	0
10	0	1	0	0	0	0	1	0
11	0	1	0	0	1	0	0	0
12	0	1	0	0	1	0	1	0
13	0	1	0	0	0	0	0	0
14	0	1	0	0	0	0	1	0
15	0	1	0	0	1	0	0	0
16	0	1	0	0	1	0	1	0
17	0	1	0	0	0	0	0	0
18	0	1	0	0	0	0	1	0
19	0	1	0	0	0	1	0	0
20	0	1	0	0	0	1	1	0
21	0	0	1	0	0	0	0	0
22	0	0	1	0	0	0	1	0
23	0	0	1	0	1	0	0	0
24	0	0	1	0	1	0	1	0
25	0	0	0	1	0	0	0	0
26	0	0	0	1	0	0	1	0
27	0	0	0	1	1	0	0	0
28	0	0	0	1	1	0	1	0
29	0	1	0	0	0	0	0	0
30	0	1	0	0	0	0	1	0
31	0	1	0	0	1	0	0	0
32	0	1	0	0	1	0	1	0

$$B = \begin{bmatrix} .03946 \\ .01739 \\ .02255 \\ .00590 \\ .00342 \\ .03393 \\ .00674 \\ .02834 \end{bmatrix}$$

χ^2 due to model = 539.59 d.f. = 7

χ^2 due to error = 16.238 d.f. = 24 ($p > .75$)

$R^2 = .971$

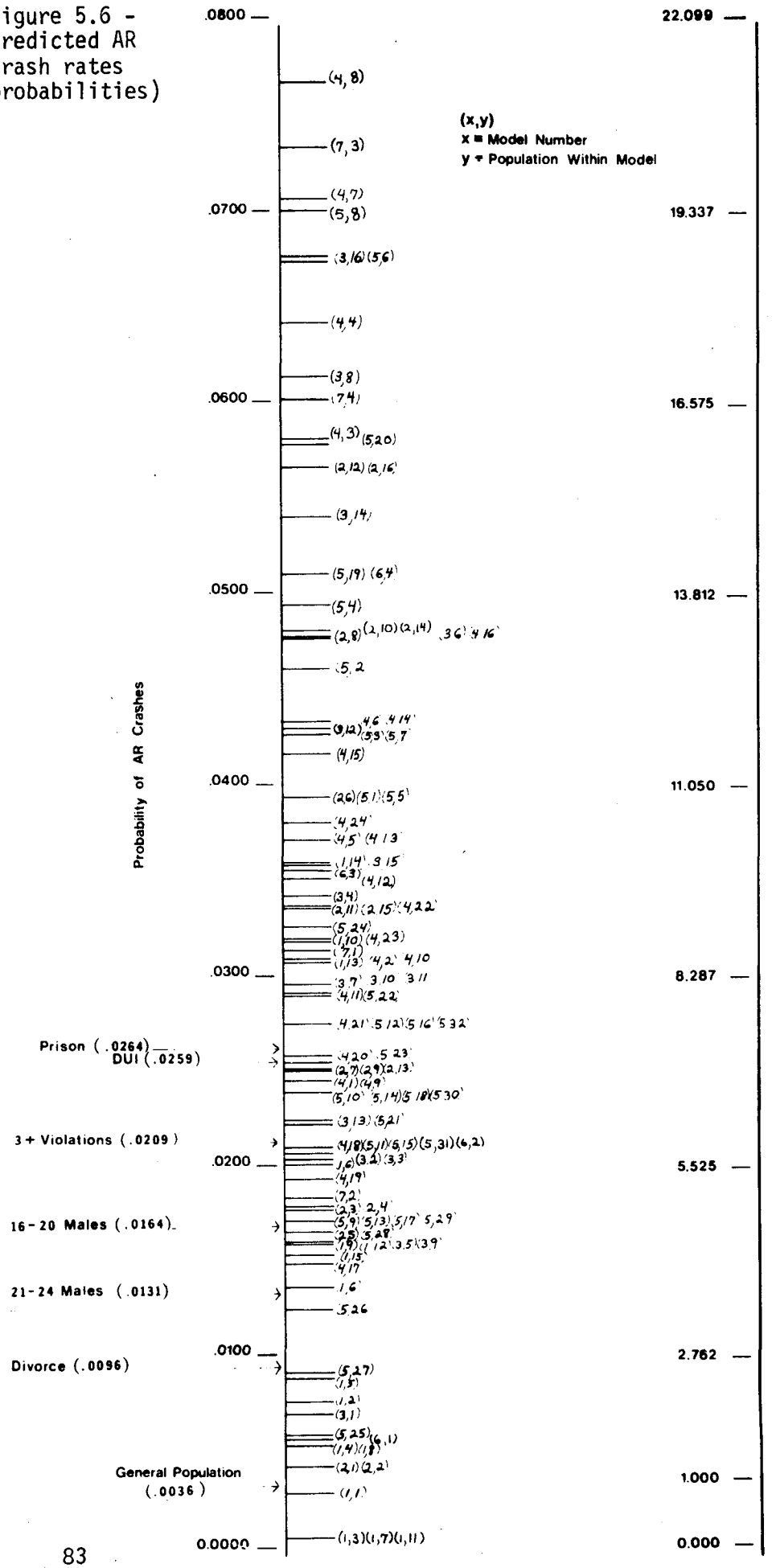
Ratio of largest predicted value to smallest = 12.09

has nearly ten times the effect that it has for all the other subpopulations. For the older males not having previous A/R crashes, the A/R crash rates are higher for drivers having some days of suspension/revocation than for those who do not. Thus, the previous A/R crash variable seems to take precedent over the total days suspension/revocation variable for the older male driver groups.

While the various subpopulations appearing in the various models are not mutually exclusive (i.e., the same drivers may be included in subpopulations of more than one high-risk group), it still may be of interest to examine the range of predicted A/R crash rates across all the high-risk groups. Figure 5.6 shows this range of predicted values graphically. The numbers in parentheses give the model number (1-7) and the subpopulation number within the model respectively. An overall A/R crash rate for the general population is shown on the chart as .00362. This rate was determined from the frequencies in the general population model and, hence, is based upon data containing complete information on all the variables used in this model and not simply A/R crash information. As would be expected, only a few subpopulations from the general population model fall below this value. On the other hand, the rates for subpopulations for other groups range up to more than twenty times this overall rate. Thus, for example, while fewer than four out of one thousand drivers selected at random from the general driving population would be expected to experience an A/R crash in one year's time, approximately seventy-seven out of one thousand selected from drivers with previous DUI's, who were under twenty years old, who had one or more speeding and one or more reckless violations in the last year, and more than 185 days of suspension/revocation in the last three years would be expected to have A/R crashes in one year's time.

In summary, the variables selected for inclusion in the A/R crash prediction models are shown to define subpopulations or subgroups of the various

Figure 5.6 - Predicted AR crash rates (probabilities)



high-risk groups which have risks of A/R crashes far exceeding those of the overall high-risk groups themselves. Categorical regression analyses produced good fitting models over the subpopulations of each high-risk group. The predicted values from these models represent smoothed (and, hence, relatively stable) estimates of the proportions of drivers expected to have A/R crashes for each subpopulation in the projection year. Tests and results of the actual predictive performance of the models are presented in the next chapter.

Chapter 6 - VALIDITY TESTING

6.1 Introduction

In order to assess the validity of the predictive models that were developed, three types of validity tests were performed. These included (1) a concurrent (or retrospective) test in which a portion of the data pool potentially usable for model development was held aside; (2) a prospective validity test in which subgroups were identified based on the chosen predictor variables (e.g., driving record data) through 1975 and their actual 1976 A/R crash rates were compared to the predicted 1976 rates; and (3) a validity test of the models' ability to predict A/R crash rates two years ahead in which subgroups were identified based on data through 1974 and their actual 1976 crash rates were compared to predicted rates. These three tests and their results are described below.

6.2 Concurrent Validity Tests

The first of these tests, termed the concurrent validity tests, was done by splitting the original data files into two parts. As noted earlier, the first part, consisting of two-thirds of the total cases, was used for the model development, while the remaining third was set aside for the purpose of concurrent validity testing. The division was done in a pseudo-random manner (i.e., every third case was set aside). Since the models were developed to fit well to the sample A/R crash proportions in the various subpopulations, the models would be expected to fit well to the remaining third of the data, provided their sample proportions were very nearly the same. Thus, in a sense, the concurrent validity tests are tests of the randomization procedure used in dividing up the data. More appropriate tests which apply the models to crash data collected in other time periods are described in Section 6.3.

In these concurrent validity tests, the A/R crash rates predicted by the models were compared to the sample proportions of the one-third control group

for each of the seven groups. A x^2 goodness of fit statistic or x^2 due to error was computed for each model using the same weighted least squares procedure as in the model development. In this case, since no parameters were being estimated, the degrees of freedom for the x^2 due to error was increased substantially. The results of the concurrent validity tests are given in Table 6.1. The table shows that for only two of the six[1] models tested is the error term highly significant--the three or more violations group and the divorce group. In both of these cases, the lack of fit stems primarily from large differences in the sample proportions of A/R crashes for a single subpopulation of the model. Specifically, in the eighth subpopulation of the three or more violations group--young males with one or more days suspension/revocation, one or more previous A/R crashes, and one or more night crashes--the actual sample A/R crash proportion for the two-thirds sample is .0809 while the one predicted by the model is .0712. However, in the one-third control sample for this group the proportion is only .0199. The difference between this value and the predicted value contributes an amount of 27.14 to the overall x^2 due to error. With this term omitted, the x^2 due to error is not significant at the 10 percent level. For the combined samples--two-thirds plus one-third--the overall sample proportion for the eighth subpopulation is .05808, which is much more in line with the predicted value.

Similarly, for the divorce model, in the fourth subpopulation--drivers with one or more previous A/R crashes and one or more reckless violations--the sample proportion is .0667 and the predicted value is .0512, while the sample proportion for the control group is .0000. Here, the subpopulation contributes

---

[1]An error in the prison group data file was discovered after new data for prospective validity tests had been collected. Since the sample size for this group was quite small, only the prospective tests were done for this group.

Table 6.1 Concurrent validity tests.

Model	χ^2	d.f.	p Value
1. General population	14.88	16	$p>.50$
2. 16-20 yr. old males	18.90	16	$p\approx.25$
3. 21-24 yr. old males	26.67	16	$p\approx.05$
4. DUI	35.35	24	$.05<p<.10$
5. Three or more violations	63.77	32	$p<.001$
6. Divorce	24.25	4	$p<.001$

nearly all of the x^2 due to error, and again, the predicted value lies between the two sample values. In the remaining 106 of the 108 subpopulations of the six models, the two sets of sample proportions were in fairly close agreement. As a result, the predicted values provide a reasonably good fit to either set of proportions. As noted earlier, however, the more important tests of true predictive validity are described below.

6.3 Prospective Validity Tests

It will be recalled that the predictive models were developed using A/R crash data for 1975 as the criterion or dependent variable and characteristics of the driver and his driving history through 1974 as predictor variables. Prospective validity tests were performed by comparing the proportions of drivers experiencing A/R crashes in 1976 with those predicted by the models under two separate conditions. The first of these constituted tests of the one-year-ahead predictive accuracy of the models--the most appropriate test in terms of the model development procedure used. For these tests the variables defined by the final models for each high-risk group were again used to define new subgroups of the high-risk groups, but this time using data on driver characteristics and driver histories through 1975. The proportion of drivers in each of these subgroups having A/R crashes in 1976 was then compared with the proportion predicted by the appropriate model.

The second prospective tests were tests of the two-years-ahead predictive accuracy of the models. For these tests, the one-third control groups were combined with the remaining two-thirds of the data and subgroups were again defined on the basis of data through 1974. The proportions predicted by the models for each subgroup were then compared to the actual crash-involved proportions in 1976.

Several types of comparisons were made on these two sets of actual and predicted proportions. As with the concurrent validity tests, the fit of the models to the two sets of actual proportions were analyzed using weighted x^2 tests. Tables 6.2 through 6.8 show, for each high-risk group, the proportions predicted by the model, the concurrent validity test proportions, and the two sets of prospective validity proportions. The actual frequencies for the prospective validity tests appear in Appendix D. Table 6.9 shows the x^2 goodness of fit tests for both the one and two year predictions for each high-risk group. This x^2 and its significance level are shown on the top line of each cell. From the table it can be seen that the lack of fit is highly significant in every case except the one-year-ahead predictions for the prison group. This indicates that for each group, the proportion of drivers having A/R crashes has changed significantly over time for at least some of the subgroups.

Obviously, in building predictive models, it would be desirable that the models accurately predict the future proportions (i.e., the proportions would not change significantly over time). Since Table 6.9 indicates that there were significant changes, the logical question that arises concerns whether or not the models are providing useful information: is relevant predictive information being gained by subdividing the high-risk groups into subgroups using the models or would a single high-risk group mean suffice. To examine this question for the one and two-year data, three other quantities were computed for each test. The first two of these (shown on the second and third lines of the cells of Table 6.9) are a x^2 due to variation about the overall group proportion, and the ratio of the x^2 due to error about the model to the x^2 due to variation about the overall group proportion. These two quantities provide an indication of how much of a variation in the proportions of drivers having A/R crashes across the subgroups of a given high-risk group is accounted for by the model

Table 6.2 General population - validity test comparisons.

	Total Days S/R	Acc. Viol.	Sex	Reckless	Predicted Proportions	Concurrent Validity Proportions	(New) Prospective 1 Year Test Proportions	(Old) Prospective 2 Year Test Proportions
1.	N	N	M	N	.0028	.0033	.0034	.0030
2.	N	N	M	S	.0077	.0102	.0109	.0105
3.	N	N	F	N	.0005	.0006	.0006	.0004
4.	N	N	F	S	.0054	.0038	.0011	.0050
5.	N	S	M	N	.0090	.0082	.0087	.0093
6.	N	S	M	S	.0139	.0238	.0190	.0101
7.	N	S	F	N	.0005	.0007	.0014	.0010
8.	N	S	F	S	.0054	.0156	.0052	.0073
9.	S	N	M	N	.0161	.0160	.0189	.0142
10.	S	N	M	S	.0319	.0298	.0282	.0211
11.	S	N	F	N	.0005	.0067	.0070	.0022
12.	S	N	F	S	.0163	.0000	.0638	.0211
13.	S	S	M	N	.0311	.0248	.0294	.0256
14.	S	S	M	S	.0360	.0332	.0400	.0400
15.	S	S	F	N	.0155	.0000	.0112	.0243
16.	S	S	F	S	.0204	.0344	.0000	.0000

Table 6.3 16-20 year old males - validity test comparisons.

	Total Days S/R	Total Viol.	Night Crashes	Night Viol. Arrests	Predicted Proportions	Concurrent Validity Proportions	Prospective 1 Year Test Proportions	Prospective 2 Year Test Proportions
1.	N	N	N	N	.0093	.0085	.0117	.0120
2.	N	N	N	S	.0093	.0384	.0628	.0547
3.	N	N	S	N	.0179	.0176	.0202	.0193
4.	N	N	S	S	.0179	.0000	.0238	.0541
5.	N	S	N	N	.0166	.0153	.0210	.0183
6.	N	S	N	S	.0396	.0533	.0531	.0362
7.	N	S	S	N	.0252	.0297	.0315	.0311
8.	N	S	S	S	.0481	.0196	.0440	.0606
9.	S	N	N	N	.0253	.0320	.0391	.0350
10.	S	N	N	S	.0482	.0000	.0000	.0000
11.	S	N	S	N	.0339	.0438	.0431	.0347
12.	S	N	S	S	.0568	.1250	.1818	.0526
13.	S	S	N	N	.0253	.0323	.0293	.0288
14.	S	S	N	S	.0482	.0583	.0541	.0505
15.	S	S	S	N	.0339	.0526	.0369	.0501
16.	S	S	S	S	.0568	.0638	.0520	.0565

Table 6.4 21-24 year old males - validity test comparisons.

	Total Days S/R	Reckless Viol.	Alcohol Viol.	A/R Crashes	Predicted Proportions	Concurrent Validity Proportions	Prospective 1 Year Test Proportions	Prospective 2 Year Test Proportions
1.	N	N	N	N	.0069	.0065	.0081	.0069
2.	N	N	N	S	.0205	.0371	.0410	.0268
3.	N	N	S	N	.0205	.0287	.0278	.0227
4.	N	N	S	S	.0340	.0000	.0256	.0513
5.	N	S	N	N	.0162	.0143	.0162	.0136
6.	N	S	N	S	.0480	.0449	.0422	.0333
7.	N	S	S	N	.0297	.0000	.0345	.0510
8.	N	S	S	S	.0616	.0000	.0455	.0370
9.	S	N	N	N	.0162	.0136	.0179	.0152
10.	S	N	N	S	.0297	.0540	.0441	.0655
11.	S	N	S	N	.0297	.0321	.0393	.0306
12.	S	N	S	S	.0433	.0454	.0447	.0479
13.	S	S	N	N	.0224	.0315	.0272	.0219
14.	S	S	N	S	.0542	.0507	.0673	.0592
15.	S	S	S	N	.0359	.0351	.0484	.0413
16.	S	S	S	S	.0678	.0282	.0626	.0626

Table 6.5 DUI group - validity test comparisons

	Age	Speeding Viol.	Days S/R	Reckless	Predicted Proportions	Concurrent Validity Proportions	Prospective 1 Year Test Proportions	Prospective 2 Year Test Proportions
1.	1	N	N	N	.0248	.0225	.0595	.0446
2.	1	N	N	S	.0309	.0526	.0833	.0545
3.	1	N	S	N	.0583	.0779	.0562	.0601
4.	1	N	S	S	.0644	.0000	.0517	.0556
5.	1	S	N	N	.0374	.0000	.0556	.0473
6.	1	S	N	S	.0435	.0000	.0526	.0714
7.	1	S	S	N	.0709	.0370	.0825	.0714
8.	1	S	S	S	.0770	.0000	.0588	.0690
9.	2	N	N	N	.0248	.0153	.0368	.0442
10.	2	N	N	S	.0309	.0491	.0331	.0468
11.	2	N	S	N	.0292	.0342	.0397	.0383
12.	2	N	S	S	.0353	.0493	.0580	.0476
13.	2	S	N	N	.0374	.0410	.0266	.0256
14.	2	S	N	S	.0435	.0434	.0159	.0380
15.	2	S	S	N	.0418	.0320	.0414	.0269
16.	2	S	S	S	.0479	.0000	.0488	.0303
17.	3	N	N	N	.0151	.0158	.0217	.0207
18.	3	N	N	S	.0212	.0297	.0301	.0254
19.	3	N	S	N	.0195	.0188	.0211	.0202
20.	3	N	S	S	.0256	.0409	.0347	.0395
21.	3	S	N	N	.0277	.0158	.0247	.0258
22.	3	S	N	S	.0338	.0000	.0367	.0147
23.	3	S	S	N	.0321	.0161	.0366	.0301
24.	3	S	S	S	.0382	.0000	.0523	.0578

Table 6.6 Three or more violations - validity test comparisons.

	Age	Sex	Days S/R	A/R Crashes	Night Crashes	Predicted Proportions	Concurrent Validity Proportions	Prospective 1 Year Test Proportions	Prospective 2 Year Test Proportions
1.	Y	M	N	N	N	.0395	.0364	.0294	.0246
2.	Y	M	N	N	S	.0462	.0167	.0391	.0300
3.	Y	M	N	S	N	.0429	.0333	.0571	.0244
4.	Y	M	N	S	S	.0496	.0000	.0528	.0444
5.	Y	M	S	N	N	.0395	.1153	.0391	.0399
6.	Y	M	S	N	S	.0678	.0000	.0319	.0446
7.	Y	M	S	S	N	.0428	.0454	.0547	.0714
8.	Y	M	S	S	S	.0712	.1999	.0643	.0622
9.	Y	F	N	N	N	.0174	.0405	.0070	.0107
10.	Y	F	N	N	S	.0241	.0101	.0238	.0067
11.	Y	F	N	S	N	.0208	.0531	.0000	.0000
12.	Y	F	N	S	S	.0278	.0000	.0000	.2143
13.	Y	F	S	N	N	.0174	.1162	.0143	.0063
14.	Y	F	S	N	S	.0241	.0000	.0213	.0345
15.	Y	F	S	S	N	.0208	.0199	.3333	.0000
16.	Y	F	S	S	S	.0276	.1999	.1250	.0000
17.	0	M	N	N	N	.0174	.0170	.0157	.0139
18.	0	M	N	N	S	.0241	.0045	.0252	.0209
19.	0	M	N	S	N	.0513	.0260	.0434	.0352
20.	0	M	N	S	S	.0581	.0131	.0535	.0372
21.	0	M	S	N	N	.0226	.0583	.0252	.0220
22.	0	M	S	N	S	.0293	.0625	.0315	.0237
23.	0	M	S	S	N	.0260	.0434	.0395	.0363
24.	0	M	S	S	S	.0327	.0392	.0424	.0427
25.	0	F	N	N	N	.0059	.0213	.0054	.0059
26.	0	F	N	N	S	.0126	.0084	.0121	.0088
27.	0	F	N	S	N	.0093	.0241	.0375	.0235
28.	0	F	N	S	S	.0161	.0227	.0417	.0327
29.	0	F	S	N	N	.0174	.0252	.0151	.0070
30.	0	F	S	N	S	.0241	.0263	.0126	.0000
31.	0	F	S	S	N	.0208	.0418	.0481	.0097
32.	0	F	S	S	S	.0276	.0338	.0000	.0000

Table 6.7 Divorce group - validity test comparisons.

	Alcohol Violations	Reckless	Predicted Proportions	Concurrent Validity Proportions	Prospective 1 Year Test Proportions	Prospective 2 Year Test Proportions
1.	N	N	.0057	.0046	.0066	.0053
2.	N	S	.0212	.0294	.0309	.0155
3.	S	N	.0357	.0222	.0234	.0189
4.	S	S	.0512	.0000	.0440	.0243

Table 6.8 Prison group - validity test comparisons.

	Admin. Viol.	Age	Predicted Proportions	Concurrent Validity Proportions	Prospective 1 Year Test Proportions	Prospective 2 Year Test Proportions
1.	N	Y	.0315		.0269	.0234
2.	N	0	.0184	not done[1]	.0217	.0198
3.	S	Y	.0734		.0641	.0357
4.	S	0	.0602		.0303	.0143

[1]Concurrent validity tests were not done due to the small sample size for this group and the fact that the perspective data was available at the time the model was completed.

Table 6.9. Goodness-of-fit tests for prospective validity proportions.

		One-Year-Ahead Tests		Two-Years-Ahead Tests	
General Population	$\chi^2\hat{p}$ (p)	58.99	(<.0005)	35.24	(.005)
	$\chi^2 p_0$	668.85		652.95	
	$\chi^2\hat{p}/\chi^2 p_0$	.090		.054	
16-20 Yr. Males	$\chi^2\hat{p}$ (p)	103.83	(<.0005)	92.97	(<.0005)
	$\chi^2 p_0$	234.87		194.84	
	$\chi^2\hat{p}/\chi^2 p_0$	.442		.477	
21-24 Yr. Males	$\chi^2\hat{p}$ (p)	63.56	(<.0005)	55.31	(<.0005)
	$\chi^2 p_0$	467.61		424.01	
	$\chi^2\hat{p}/\chi^2 p_0$	.136		.130	
DUI	$\chi^2\hat{p}$ (p)	64.25	(<.0005)	66.13	(<.0005)
	$\chi^2 p_0$	43.89		46.24	
	$\chi^2\hat{p}/\chi^2 p_0$	1.464		1.430	
3+ Violations	$\chi^2\hat{p}$ (p)	155.22	(<.0005)	394.75	(<.0005)
	$\chi^2 p_0$	619.67		848.83	
	$\chi^2\hat{p}/\chi^2 p_0$	.251		.465	
Divorce	$\chi^2\hat{p}$ (p)	9.93	(<.025)	23.69	(<.0005)
	$\chi^2 p_0$	20.72		17.57	
	$\chi^2\hat{p}/\chi^2 p_0$	.479		1.348	
Prison	$\chi^2\hat{p}$ (p)	2.08	(>.10)	16.33	(<.0005)
	$\chi^2 p_0$	1.26		1.06	
	$\chi^2\hat{p}/\chi^2 p_0$	1.651		15.406	

$\chi^2\hat{p} \equiv$ Chi square due to error about model, $\chi^2 p_0 \equiv$ Chi square due to error about overall proportion

in comparison to simply replacing the predicted subgroup proportions by the overall group proportion. Thus, if the model is providing more information than the simple overall group proportion, the ratio will be quite small (i.e., the actual data points will vary about the overall proportion a great deal more than they will vary about the individual subgroup proportions predicted by the models). In general, the two lower numbers in each cell indicate this to be the case for the one-year-ahead tests. For example, for the general population, the variation about the model-predicted subgroup proportions is only one-tenth of the variation about the overall group proportion, indicating that the model, although indicating a significant lack of fit, provides a large amount of useful predictive information. Specifically, for both of the tests which involve the general population, for both of the tests which involve the 21 to 24-year-old male group, and for the one-year-ahead tests of the 3+ violation group, the models account for, by far, the major part of the variation in the A/R crash proportions across the subgroups. The models also seem to do moderately well in this regard for the 16 to 20-year-old males, the two-years-ahead predictions for the 3+ violation group, and the one-year-ahead predictions for the divorce group. They do not do very well by this criterion for the DUI group, nor for the two-years-ahead predictions for the divorce group and for both predictions for the prison group.

To further examine the significant lack of fit indicated by the model, individual cell contributions to the overall chi-square for lack of fit were examined to see if patterns existed. For example, it would be of interest to know whether or not the cells (subgroups) which contributed the largest values to the significant chi-square were cells which would be important in real world use of the models--the higher probability subgroups. This subgroup by subgroup examination was done for each of the models for the one year prospective test. The results were consistent and somewhat encouraging. For example, for the

16 to 20-year-old males, it was noted that the single major contributor to the significant chi-square (44.9 of the 58.99 total) was subgroup 1, representing young males with no days under suspension, no alcohol violations, no night crashes and no night violations. This particular subgroup has a large sample size and a low predicted proportion of A/R crash involved drivers. Four other subgroups in the 16-20 year old males model also contribute heavily to the significant chi-square and in each case these were subgroups with large sample sizes and low to medium predicted probabilities of future crashes: subgroups which may not be as important in the actual use of the model as subgroups with a higher predicted probability of an A/R crash.

Similar examinations were carried out for each of the other models except the general population model. For the 21 to 24-year-old male group, five subgroups contributed large amounts to the chi-square. Each of the five was a large subgroup with a medium predicted probability of future A/R crash. The most significant contributor was again subgroup 1 representing the "cleanest" drivers with the lowest predicted probability of a crash. For the DUI group, the lack of fit was more evenly distributed throughout all of the subgroups. However, the three subgroups contributing most to the significant chi-square were subgroups with large sample sizes and low predicted probabilities. For the 3+ plus violations group consisting of 32 subgroups, eight subgroups contributed large amounts to the total chi-square. The pattern here was the same as before with these subgroups having large sample sizes and low predicted proportions of crash involvement. The one exception in this subgroup was subgroup 6, representing three or more violation drivers who are young males with some suspensions, no alcohol-related crashes, and some night crashes, for which the model predicted a high probability of A/R crash (.0678). The actual prospective one year proportion was .0319. For the divorce group, the third subgroup accounted for the largest share of the significant chi-square. The subgroup,

representing divorcees, with some alcohol-related accidents and no reckless violations was a subgroup which contained a large sample size and was predicted to have a medium probability of A/R crash involvement when compared to the other subgroups. In actuality, this subgroup had a lower probability of crash than did most of the other subgroups. Finally, as indicated in Table 6.9, there was no significant lack of fit for the prison group in the one-year-ahead test while there was significant lack of fit in the two-years-ahead test.

In summary, the results of these cell by cell examinations tend to indicate that the heaviest contributors to the chi-squares indicating significant lack of fit for the models were subgroups which, in general, had large sample sizes (which would be expected since the sample size is a strong determinant of significance) and low to medium predicted probabilities of future A/R crash.

Several comments should be made relative to these results. First, it was noted in the chapter on model fitting that there was relatively little variation across the subpopulations of the DUI group. It is, therefore, not too surprising that the overall group proportion provides a relatively good fit to the data for this group. Second, the models were developed to be one-year-ahead predictors, so that again it is not surprising that for some groups they do not do very well as two-years-ahead predictors. Finally, it should be noted that while the lack of fit test of the model for one-year-ahead predictions for the prison group was not significant, the overall group proportion still provided a better fit to the data than did the model.

As a third criterion for testing the model predictions, rank correlations were computed between the predicted and actual proportions for each subgroup for both the one-year and the two-year predictions. These quantities tend to indicate how the ordering of the predicted proportions (from relatively low proportions of A/R crashes to relatively high proportions of A/R crashes) tend to remain stable over time. These quantities are shown in Table 6.10. Here it

Table 6.10 Rank correlations of actual and predicted proportions.

Group	One-Year Ahead Predictions		Two-Years Ahead Predictions	
	r (Spearman)	p	r (Spearman)	p
General population	.683	<.01	.731	<.01
16-20 yr. old males	.456	<.05	.355	>.05
21-24 yr. old males	.857	<.01	.762	<.01
DUI	.483	.01	.607	<.01
3+ violations	.515	<.01	.636	<.01
Divorce	.800	>.05	1.000	.05
Prison	1.000	.05	.400	>.05

is seen that in all cases except for the one-year-ahead prediction for the divorce group and for the two-years-ahead predictions for the 16 to 20-year-old male and prison groups, the rank correlations are positive and statistically significant. These results also indicate that the models give the best prediction for the general population, and the 21 to 24-year-old male groups.

Though one would prefer to have the prospective tests of individual subgroup proportions be nonsignificant along with positive and significant test results of the subgroup rank correlations, having the latter alone may be adequate from a practical standpoint. The primary objective of the project is to identify those subgroups which are at the highest risk of A/R crash involvement so that they may potentially be brought into countermeasure programs before the A/R crashes take place. The positive and statistically significant rank correlations indicate that the appropriate high-risk subgroups have been identified for such action. The major drawback would be in terms of doing an a-priori cost effectiveness analysis as described in Chapter 4 of Volume II. In this type of application of the models, the predicted proportion of A/R crashes for a particular subgroup is used to help assign an anticipated benefit from a countermeasure program in terms of the costs saved by reducing the number of crashes in that group. This type of economic analysis is sensitive to fluctuations in predicted proportions. If the actual crash experience is different from the one that was predicted, the potential payoff will also be different. However, the ranking of various countermeasure programs for a particular subgroup will remain, for the most part, constant.

Of course, the predicted proportion for a particular subgroup should never be used as a substitute for an actual control group in the evaluation of any countermeasure program. This issue is discussed in some detail in Chapter 5 of Volume II.

Since many of the high-risk subpopulations identified in the models are quite small, in practice it may be the case that the highest risk subgroups for some high-risk groups will be grouped for countermeasure application. In addition, it is fair to say that real world use of these models will be concentrated in the higher risk subgroups. That is, if a countermeasure is to be implemented, it will be applied to the higher risk subgroups in a given high-risk group rather than to the lower risk subgroups.

Because of these facts, it is of interest to compare the actual and predicted crash rates for the classes of drivers so identified. For example, suppose the worst (in the sense of having the highest A/R crash rate) 25 percent of all subgroups for each group were to be selected for some action. Table 6.11 shows the predicted and actual percentages of drivers having A/R crashes of these "worst" subgroups for each group, based on the one-year-ahead predictions. For most groups it can be seen that the actual and predicted rates agree quite well. For all except the general population group, the actual and the predicted percentages are both about 5 percent, which is two to four times higher than the overall group percentages. Thus, by using the predicted value to identify the "worst" subgroups, a class of drivers would be identified having A/R crash rates two to four times higher than those for the group as a whole. The 25 percent figure was completely arbitrary, and in practice one might choose 10 percent, 50 percent, etc.

In summary, the prospective validity tests show that for at least some of the subgroups the proportion of drivers having A/R crashes predicted by the models changed significantly over time, causing the lack of fit test to be statistically significant in virtually all cases. On the other hand, for some groups the models still accounted for most of the variation across the subpopulation of the group. Finally, the rank correlations showed that the relative rankings of the subpopulations tended to remain stable over time.

Table 6.11. Observed and predicted A/R crash rates for the top 25% of all subgroups for each high risk group.

Group	Predicted Percentage	Observed Percentage	Overall Group Percentage
General population	3.25	3.14	0.36
16-20 yr. old males	5.05	5.38	1.64
21-24 yr. old males	5.57	5.62	1.31
DUI[1]	5.79	5.54	2.59
3+ violations	5.58	4.82	2.09
Divorce[2]	5.12	4.40	0.96
Prison[2]	7.34	6.41	2.64

[1]Seven of the twenty-four subgroups were included since two subgroups had the same 6th highest A/R crash percentage.

[2]For the divorce and prison groups the predicted and actual percentages are for a single subgroup.

In addition, it appears that the models more accurately predict the actual proportion of crash-involved drivers in the higher risk subgroups--the subgroups which would be of greatest interest to a program administrator.

CHAPTER 7 - CONCLUSIONS AND RECOMMENDATIONS

7.1 Introduction

The two major questions addressed by the project were: (1) Can individuals at a high risk of alcohol-related crash involvement be identified before they have the A/R crash? and (2) Can effective countermeasures which are appropriate to such high-risk individuals be identified from currently available information? As shown in Chapters 5 and 6 of this volume, individuals at an elevated risk of alcohol-related crash involvement can be identified. In fact, groups of individuals with risks as much as 20 times that of the general driving population and somewhat larger groups with risks nearly 15 times greater than average were identified using the modelling procedures adopted for this project. Variables, such as days under suspension and revocation, age, and reckless driving violations consistently appeared as predictive variables. However, countermeasures that are demonstrably effective in markedly reducing the likelihood of A/R crash involvement for the identified individuals do not exist in the current literature (the countermeasure review, which is covered in Chapter 3 of Volume II, is also summarized in this chapter). The potential utility of the models in an operational framework and the implications of the results of this project for future research are also discussed below.

7.2 A/R Crash Prediction

The prospective tests of the predictive validity of the models which are discussed in Chapter 6 assess in some detail the effectiveness of the models developed under this project. Basically, they indicate that though the models do not accurately predict the exact proportion of the individuals within each high-risk subgroup which will be involved in an A/R crash, they do identify a set of subgroups within each high-risk group that are at the highest risk of

A/R crash involvement. A/R crashes are inherently low probability events as evidenced by the fact that less than four-tenths of one percent of the general driving population are involved in such crashes in a given year. Thus, even the highest risk subgroups have predicted probabilities of A/R crash involvement of less than .10.

From an operational standpoint, assuming that one plans to implement a countermeasure program, the models identify the highest risk subgroups for program inclusion. The major drawback is not being able to predict exactly the proportion of individuals within the subgroup which will be A/R crash-involved. If one is ranking countermeasure approaches (see Volume II, Chapter 4) for application to a particular group (e.g., DUI), this should not pose a problem because the difference in predicted and actual proportions should similarly affect all the countermeasure approaches under consideration. However, if one were considering countermeasure approaches in terms of which high-risk groups to address, deviations in the actual A/R crash experience from the predicted proportions could influence countermeasure rankings. Nonetheless, in all likelihood, the differences would be relatively small, and high-risk subgroups would have been accurately targeted for countermeasure activity.

7.3 Acceptability of the Models

7.3.1 Practical considerations of acceptability.

One consideration in developing predictive tools to be used to identify persons for countermeasure activity is whether the models have an intuitively satisfactory rationale as well as a statistical one. In this case the variables which are used to identify the high-risk individuals should ideally appear to the layman to be related to crash behavior. In other words, a judge is much more likely to agree to put a 16-year-old male into an alcohol countermeasure

program on the basis of having a nighttime crash and several violations than for having a series of unpaid parking tickets and not reporting a change of address. Likewise, the person being placed in the countermeasure program may be more receptive to the intervention if the rationale is self-evident. All of the variables which entered into the predictive equations do carry this intuitive as well as statistical relationship to alcohol-related crashes.

Another consideration is whether the variables used are controversial or would involve a perceived invasion of privacy. The only variables used that are not directly driving-related are age, sex, divorce, and prior imprisonment. There is a long history of considering age and sex in conjunction with driving risk as evidenced by probationary licensing of young persons and differential liability insurance rates for young males.

The facts of a previous divorce or imprisonment are both part of the public record, but could conceivably be somewhat more controversial. However, little argument would probably be given to cautioning recent divorcees about a high A/R crash risk on the basis of their also having alcohol violations and reckless driving convictions. In fact, a possible scenario for the use of the models for this group might be for a judge to recommend some level of counseling to a recently divorced person with prior alcohol violations upon sentencing for a reckless driving conviction. With the current emphasis on giving ex-prisoners a new lease on life, using prior incarceration as part of the basis for any but the least threatening of countermeasure activities (i.e., warning letters) could be problematic.

In actual application, the point of intervention is another issue of importance. Conceivably, one could apply the models to a given driving population, identify individuals falling into high-risk subgroups and initiate a countermeasure activity. However, the most likely application would be to

monitor the driving records of high-risk groups and intervene as individuals committed acts (e.g., a nighttime crash) which brought them into a high-risk subgroup.

Another consideration is whether the models are simple enough to be readily applied in an operational context. Using the approach of first identifying high-risk groups and then developing a model for each one led to reasonably straightforward models. The most complex model (3+ violations) uses only five variables in addition to the initial variable describing the high-risk group. Each of the models for high-risk groups yielded high-risk subgroups with much higher predicted probabilities of A/R crash involvement than the general population model. Conceivably, a single general population model could be developed which uniquely describes each of the highest risk subgroups identified in each of the separate models. However, such a model would be extraordinarily complicated and cumbersome.

7.3.2 False positives and false negatives.

Another concern in using a predictive technique as a tool in identifying persons for countermeasure activities is the extent to which one may be treating individuals who were not going to have a crash regardless of the countermeasure (false positives) and, conversely, are not treating persons who will have a crash (false negatives). In the approach taken in this project, groups of individuals are assigned probabilities of subsequent A/R crash involvement ranging from very close to zero to a high of .07701. So, in essence, the output of the models is not that one individual is going to have an A/R crash and another is not but rather that one is more likely to have a crash than another. Because A/R crashes themselves are so rare, a person with a high probability of A/R crash involvement in relative terms has what would be perceived by the public to be a low actual probability.

Thus, in the example above, if the model were completely accurate in its predictions for the subsequent year, for even the highest risk subgroup, 94 out of 100 persons would not have an A/R crash. It becomes apparent that the concept of false positives and negatives is not appropriate in this context. What is appropriate is to consider whether the models reliably identify those subgroups which are at the highest and lowest risk of A/R crash involvement so that those groups of individuals most in need of countermeasure activity are not ignored and those who need it the least are not unnecessarily inconvenienced. The rank correlation analysis reported in Chapter 6 and summarized in Table 6.10 indicates that the models do perform reasonably well in this respect.

Ideally, one would like to identify with certainty those who would and would not have an A/R crash in a particular time frame. However, to accomplish this, in a prospective sense, even with costly and difficult to obtain psychological and social profiles, is a virtual impossibility.

7.4 Effect of Modelling Group Size on Potential Impact

As mentioned in Chapter 4 and depicted in Figure 4.1, the variable selection process in model development resulted in substantial reduction in group size and consequently in potential impact on the total A/R crash problem.

For example, though all 16 to 20-year-old males accounted for 17.31 percent of all 1975 A/R crashes, the total group identified with the necessary variables to prospectively test the model on 1976 A/R crashes accounted for only 11.44 percent of the crashes. In terms of a risk index, the newly identified group had a population based rate of A/R crashes 4.21 times greater than the general population sample. Table 7.1 presents these figures for all high-risk groups.

Table 7.1. 1975 impact and risk indices of high-risk groups identified for one-year prospective test.

Group	Impact index	Risk index
16-20 year old males	11.44	4.21
21-24 year old males	10.58	3.36
DUI	6.96	6.64
3+ violations	17.72	5.36
Divorce	.65	2.47
Prison	.21	6.78

If one assumed that only the top 25 percent of the subgroups in terms of high A/R crash risk were likely to be addressed by a countermeasure program, the percentage of all A/R crashes potentially impacted by the countermeasure program would be considerably less than the 11.4 percent total accounted for by the 16 to 20-year-old male group, for example. The top four groups in the 16 to 20-year-old male group accounted for .37 percent of all 1976 A/R crashes in North Carolina. However, these four subgroups collectively had a risk of A/R crashes 13.79 times higher than the general population. Table 7.2 presents these figures for the six high-risk groups.

Table 7.2. Impact and risk indices for the 25 percent highest risk subgroups of the high-risk groups identified for the one-year prospective validity test.

Group	Impact index	Risk index
16-20 year old males	.37	13.79
21-24 year old males	.56	14.43
DUI*	.31	14.84
3+ violations	1.81	12.37
Divorce	.04	11.27
Prison	.03	16.44

*21 percent highest risk subgroups for this high-risk group.

It is obvious that the impact index figures presented in Table 7.2 do indicate that countermeasure activities directed at the highest risk subgroups identified by the models cannot be expected to result in a large percentage reduction in A/R crashes as a whole. However, the models may be useful in targeting countermeasure programs. Many programs, by their very nature, cannot be directed at more than very small segments of the driving population. Though some may be relatively expensive on an individual basis, by directing them at those individuals most likely to need their potential effect, they can be applied more efficiently and cost effectively. This issue is discussed in some detail in Chapter 4 of the User Manual.

7.5 Countermeasure Effectiveness Levels

A major objective of the project was to identify from the literature effective countermeasures that could be used in conjunction with the models in accomplishing a reduction of A/R crashes. The results of that literature search, which are fully presented in Chapter 3 of Volume II, are summarized here.

Since the focus of the project was on identifying individuals or small groups of individuals at high-risk of A/R crashes, the countermeasure review was focused on countermeasures that would be appropriate to an individual or small group setting. Thus, countermeasures such as public information and education programs and increased enforcement were not considered in this review. A problem similar to that encountered in the high-risk group selection literature review was also present here. That is, many countermeasures were evaluated in terms other than A/R crash reduction. Most evaluations were done in terms of reduced DUI recidivism, overall accident experience, or violation experience. Another difficulty was that very few of the evaluations were conducted using a

fundamentally sound experimental design and, thus, the results of such evaluations had to be viewed with some measure of caution.

Twelve different types of alcohol countermeasures are briefly described in Volume II, a summary of their evaluations presented along with estimates of their effectiveness in reducing A/R crashes, their implementation costs, and their duration of effectiveness. Table 7.3 presents this information and also appears as Table 3.1 in Volume II. Estimates of cost and period of effectiveness are provided as inputs to the economic analysis program for use in applying cost effectiveness procedures in countermeasure selection. It should be emphasized that the estimates of effectiveness in A/R crash reduction and of periods of effectiveness are estimates based on the best available information, which, in many cases, is very limited. The initial intent was to present level of effectiveness by high-risk group. However, in nearly every case, evaluations were done in terms of all drivers rather than specific subgroups.

7.6 Summary Conclusions

Based on the experience of this project, certain conclusions can be drawn:

1. Predictive models can be developed using information available to alcohol and driver program administrators.

2. A benefit accrues in terms of higher predicted A/R crash probabilities by developing several models for individual high-risk groups over developing just one model for the general population as a whole.

3. The models are reasonably reliable predictors of alcohol-related crash experience in terms of ranking subgroups by risk, even when tested in a prospective sense of predicting a one-year crash experience.

Table 7.3. Summary of countermeasure information.

Countermeasure	Estimated A/R Crash Reduction Potentials			Treatment Costs	Estimated Effectiveness Period
	A/R Fatal Crashes	A/R Injury Crashes	A/R PDO Crashes		
1. Warning letters	4-20%	4-20%	4-20%	$ 0.50 - $ 1.50/letter	5 mo. - 7 mo.
2. Driver improvement clinics	0-?	0-?	0-?	$ 10.00 - $ 30.00/driver	?
3. Hearings	0-15%	0-15%	0-15%	$ 20.00 - $ 70.00/driver	7 mo. - 1 year
4. Probationary license for DUI first offenders	6-12%	6-12%	6-12%	Depends on information provided to judges	1 year
5. Short-term rehabilitation programs - Type 1-3 alcohol safety schools	?	?	?	$ 10.00 - $ 70.00/driver depending on school type	6 months?
6. Short-term rehabilitation programs - Power Motivation Training	?	?	?	$ 75.00 - $100.00/driver	?
7. Suspension/revocation of license	25-35%	25-35%	25-35%	Unknown, but relatively low	2 yrs. - 4 yrs.
8. Group therapy[ma]	?	?	?	$ 3.00 - $ 10.00/subject/session	?
9. Alcoholics Anonymous[a]	2-40%	2-40%	2-40%	?	?
10. Psychotherapy[a]	0-?	0-?	0-?	?	?
11. Aversion therapy[a]	10-40%	10-40%	10-40%	?	1-2 years
12. Direct chemotherapy[a]	2-50%	2-50%	2-50%	$150.00 - $200.00/driver/year	Throughout continued treatment

ma - most suitable for mid-range problem drinkers and alcoholics
a - suitable only for alcoholics within high-risk groups

4. A/R crashes are such low probability events in the general driving population that, even when a person is identified with a risk as much as 20 times greater than average, the probability of A/R crash involvement in the next year is still less than .08.

5. Potential cost effectiveness for countermeasure programs with limited A/R crash reduction ability can be demonstrated by applying them to the high-risk subgroups identified in the models.

6. Currently, few sound evaluations documenting the true traffic safety benefits of alcohol countermeasures are available.

7.7 Recommendations

Based on the experiences of conducting this project, certain recommendations are forthcoming.

1. In the application of the models, groupings of the highest-risk subgroups (e.g., the 25 percent highest risk subgroups) within a high-risk group should be used for countermeasure implementation. This increases the number of individuals affected over using just the single highest risk subgroup.

2. Consideration should be given to developing models designed to be predictive of two year A/R crash probabilities. It is likely that by using this approach, high-risk subgroups could be identified with predictor probabilities higher than those attained for the one-year period. A period longer than two years is probably not advisable because countermeasures which might be applied as a result of the modelling outputs characteristically do not have an estimated period of effectiveness longer than two years.

3. The current effort to conduct scientifically sound evaluations of specific alcohol countermeasures should be continued and expanded.

4. The tools developed under this project as presented in Volume II, should be used to assist in targeting populations for countermeasure implementation and monitoring countermeasure effectiveness.

REFERENCES

Borkenstein, R.F., Crowther, R.F., Shumate, R.P., Zeil, W.B., & Zylman, R. The role of the drinking driver in traffic accidents. *Bloomington:* Indiana University Department of Police Administration, 1964.

Burg, A. Vision test scores and driving record: Additional findings. Los Angeles: UCLA Institute of Transportation and Traffic Engineering, 1968. [Report No. 68-27]

Cahalan, D. Problem drinkers. San Francisco: Jassey-Bass, Inc., 1970.

Cahalan, D., & Room, R. Problem drinking among American men. New Brunswick, NJ: Rutgers Center of Alcohol Studies, 1974.

Carlson, W.L. Alcohol usage of the nighttime driver. Journal of Safety Research, 1972, 4(1), 12-25.

Carlson, W.L. Age, exposure, and alcohol involvement in night crashes. Journal of Safety Research, 1973, 5, 247-259.

Clarke, S., & Koch, G.G. What determines whether persons arrested for burglary and larceny go to prison? Chapel Hill: University of North Caroline Department of Biostatistics, 1974.

Filkins, L.D., Compton, C.P., Douglass, R.L., & Flora, J.D. Analysis of high risk groups for alcohol countermeasures. Ann Arbor: University of Michigan Highway Safety Research Institute, 1975. [DOT-HS-801-434]

Filkins, L.D., Clark, C.D., Rosenblatt, C.A., Carlson, W.L., Kerlan, M.W., & Hanson, H. Alcohol abuse and traffic safety: A study of fatalities, DWI offenders, alcoholics, and court-related treatment approaches. Ann Arbor: University of Michigan Highway Safety Research Institute, 1970. [DOT-HS-800-409]

Fleiss, J.L. Statistical methods for rates and proportions. New York: John Wiley and Sons, 1973.

Goldstein, L.G. Youthful drivers as a special safety problem. In P.F. Waller, Ed., The young driver: Reckless or unprepared? Chapel Hill: University of North Carolina Highway Safety Research Center, 1971.

Grizzle, J.E., Starmer, C.F., & Koch, G.G. Analysis of categorical data by linear models. Biometrics, 1969, 25(3), 489-504.

Guze, S.B., Tuason, V.B., Gatfield, P.D., Stewart, M.A., &Pichen, B. Psychiatric illness and crime with particular reference to alcoholism: A study of 223 criminals. Journal of Nervous and Mental Disease, 1962, 134(6), 512-521.

Harano, R.M. The psychometric prediction of negligent driver recidivism. Sacramento: State of California Department of Motor Vehicles Research and Statistics Section, 1974.

Harano, R.M., McBride, R.S., & Peck, R.C. The prediction of accident liability through biographical data and psychometric tests. Sacramento: State of California Department of Motor Vehicles Research and Statistics Section, 1973.

Holcomb, R.L. Alcohol in relation to traffic accidents. Journal of the American Medical Association, 1938, 111(12), 1076-1085.

Indiana University Institute for Research in Public Safety. Tri-level study of the causes of traffic accidents. Bloomington: Author, 1973.

Kephart, W.M. Drinking and marital disruption. Quarterly Journal of Studies on Alcohol, 1954, 15(1), 63-73.

Levinger, G. Sources of marital dissatisfaction among applicants for divorce. American Journal of Orthopsychiatry, 1966, 36(5), 803-807.

Li, L.K., & Waller, P.F. Evaluation of the North Carolina habitual offender law. Chapel Hill: University of North Carolina Highway Safety Research Center, 1976.

Marden, P.G., & Kolodner, K. Alcohol use and abuse among adolescents. (NCAI026533) Canton, NY: St. Lawrence University, n.d.

Marsh, W.C., & Hubert, D.M. The prediction of driving record following driver improvement contacts. Sacramento: State of California Department of Motor Vehicles Statistics and Research Section, 1974.

McMurray, L. Emotional stress and driving performance: The effect of divorce. Olympia: State of Washington Department of Motor Vehicles, 1968.

Minnesota Department of Public Safety. The alcohol-impaired driver and highway crashes. St. Paul: Author, 1970.

Monaco, J.P. The collection of national trend data on alcohol related crashes for comparison with alcohol safety action project results. Menlo Park, CA: SRI International, 1975-1977. 2 Volumes.

North Carolina Department of Correction. State correction statistical abstract: January through December 1974. Raleigh: Author, 1975.

O'Day, J. Drinking involvement and age of young drivers in fatal accidents. HSRI HIT Lab Reports, October 1970.

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

Pelz, D.C., & Schuman, S.H. Exposure factors in accidents and violations of young drivers. Ann Arbor: University of Michigan Highway Safety Research Institute, 1971.

Pollack, Seymour, Didenko, McEachern, & Berger. Drinking driver and traffic safety project. Los Angeles: University of Southern California Public Systems Research Institute, 1972. [DOT-HS-800-699]

Rachel, J.V., Williams, J.R., Brehin, M.L., Cavanaugh, B., Moore, R.P., & Eckerman, W.C. A national study of adolescent drinking behavior, attitudes, and correlates. Durham, NC: Research Triangle Institute Center for the Study of Social Behavior, 1975.

Rosenberg, N., Laessig, R.H., & Rawlings, R.R. Alcohol, age, and fatal traffic accidents. Quarterly Journal of Studies on Alcohol, 1974, 35, 473-489.

U.S. Department of Justice. Census of state correctional facilities: 1974 advance report, cited in M. Aarens, T. Cameron, J. Roizen, R. Roizen, R. Room, D. Schneberk, & Wingard, D. Alcohol casualties and crime. Berkeley: University of California School of Public Health, 1977.

U.S. Department of Transportation. 1968 alcohol and highway safety report. Washington, D.C.: Author, 1969.

U.S. Department of Transportation. Alcohol safety action projects: Evaluation of operations--1974. Vol. 2. Detailed analysis. Washington, D.C.: Author, 1976.

Ulmer, R.G., & Preusser, D.F. Nassau County alcohol safety action project analysis of blood alcohol levels in fatally injured drivers. Darien, CT: Dunlap and Associates, Inc., 1973.

Waller, P.F., Ed. Proceedings of the North Carolina Symposium on Highway Safety. Vol. 5. The young driver: Reckless or unprepared? Chapel Hill: University of North Carolina Highway Safety Research Center, 1971.

Williams (1958), cited on p. A-7 in M.H. Wagner, J.H. Bigelow, J. Cobb, L. Goldstein, & R.E. Kirkpatrick, Analysis of high risk groups for alcohol countermeasures. Phase I. High risk driver study plan report. Fairfax, VA: Technical Research Associates, Inc., 1975.

Zador, P. Statistical evaluation of the effectiveness of alcohol safety action projects. Accident Analysis and Prevention, 1976, 8(1), 51-66.

Zylman, R. Youth alcohol and collision involvement. Journal of Safety Research, 1973, 5, 58-72.

APPENDIX A

Alcohol Model Study Record Format

APPENDIX A

Alcohol Model Study Record Format
(Revised May 1976)

Position	Contents
1-8	Driver License Number
9-15	Group Numbers Example: 1011001 Connotes record to be in group 1, 3, 4 & 7 1 divorce 2 prison 3 DUI 4 general population 5 16-20 males 6 21-24 males 7 3+ violations
16	Divorce 1 no divorce 2 defendant 3 plantiff
17	Prison 1 yes 2 no
18-19	Total Crashes Last 3 Years
20-21	Total A/R Crashes Last 3 Years
22-23	Total Night Crashes Last 3 Years
24-29	Date of Most Recent A/R crash (year, month,day)
30-31	Days Under Analysis
32	Control Group Number 1 Study, A/R* 2 Study, Not A/R 3 Control, A/R 4 Control, Not A/R

*A record is characterized as 'A/R' if the date of the most recent A/R crash is 1975.

Position	Contents
33-34	Total Crashes, 1973-1974
35-36	Total A/R Crashes, 1973-1974
37-38	Total Night Crashes, 1973-1974
39-41	Age of Subject

Accident Information

42 — Number of Accidents Recorded

1-3

1st. Accident

43 — Accident Year

will be '9' if accident not coded

44-45 — Accident Month

01 January
02 February
12 December
13 Not stated

46-47 — Accident Day of Month

01-31
32 Not stated

48 — Day of the Week

1 Monday
2 Tuesday
3 Wednesday
4 Thursday
5 Friday
6 Saturday
7 Sunday
8 Not stated

Position	Contents
49-52	Time of Day (24 hour clock including minutes) 0000 Midnight 1200 Noon 2460 Not stated example: 1630 - 4:30 p.m.
53	Locality 1 Business 2 Residential 3 School or playground 4 Open country (interstate or rural) 5 Not stated
54	Light Condition 1 Daylight 2 Dusk 3 Dawn 4 Darkness (street lighted) 5 Darkness (street not lighted) 6 Not stated
55	Weather 1 Clear 2 Cloudy 3 Raining 4 Snowing 5 Fog 6 Sleet or hail 7 Not stated
56	Severity (Most severe injury in accident) 1 Fatal 2 A or B class injury 3 C class injury 4 Property damage only 5 Not stated
57-58	Accident Type 01 Ran off road - right 02 Ran off road - left 03 Ran off road - straight ahead 04 Non-collision in road - overturn 05 Non-collision in road - other

Position	Contents
57-58	Accident Type (Cont')
	06 Collision of motor vehicle with pedestrian
	07 Collision of motor vehicle with parked vehicle
	08 Collision of motor vehicle with train
	09 Collision of motor vehicle with bicycle
	10 Collision of motor vehicle with animal
	11 Collision of motor vehicle with fixed object
	12 Collision of motor vehicle with other object
	13 Collision of MV with another MVs rear end - stopping or slowing
	14 Collision of MV with another MVs rear end - turning
	15 Collision of MV with another MV turning left from same roadway
	16 Collision of MV with another MV turning left across traffic
	17 Collision of MV with another MV turning right from same roadway
	18 Collision of MV with another MV turning right across traffic
	19 Collision of MV with another MV head on
	20 Collision of MV with another MV sideswipe
	21 Collision of MV with another MV at an angle
	22 Collision of MV with another MV backing
	23 Not stated
59	Total Occupants
	0-8
	9 More than 8 occupants
	- Not stated
60	Armed Forces Driver & Vehicle
	0 AF driver of unspecified vehicle
	1 AF driver of military vehicle
	2 AF driver of emergency vehicle
	3 AF driver of state owned vehicle

Position	Contents
60	Armed Forces Driver & Vehicle (Cont')
	4 AF driver of other public vehicle 5 Non AF driver of military vehicle 6 Non AF driver of emergency vehicle 7 Non AF driver of state owned vehicle 8 Non AF driver of other public vehicle 9 Not stated
61	Restriction Code
	0 None 1 Corrective lenses 2 45 mph speed limit 3 Daylight driving only 4 Corrective lenses, 45 mph speed limit and daylight driving only 5 Corrective lenses & 45 mph speed limit 6 Corrective lenses & daylight driving only 7 45 mph speed limit & daylight driving only 8 Property only 9 Other (i.e., handicaps & other) - Not applicable or not stated
62	Physical Condition
	1 Ill 2 Fatigued 3 Asleep 4 Other physical impairment 5 Restriction not complied with 6 Normal 7 Not stated
63	Sobriety
	1 Had not been drinking 2 Drinking--ability impaired 3 Drinking--unable to determine impairment 4 Not stated
64	Chemical Test
	1 Yes 2 No 3 Not stated

Position	Contents
65	Driver Charged 1 Yes 2 No 3 Not stated
66-67	Violation #1 01 Speeding below 65 mph 02 Speeding 65 to 75 mph 03 Speeding over 75 mph 04 Failed to yield right-of-way 05 Driving on wrong side of the road 06 Improper overtaking 07 Disregarded stop sign or signal 08 Disregarded traffic signal 09 Followed too closely 10 Improper turn 11 Improper or no signal 12 Improper parking location 13 Under influence of alcohol 14 Reckless driving 15 Racing 16 Failed to see if movement could be made in safety 17 Passed on curve 18 Passed on hill 19 Passed stopped school bus 20 Improper lights 21 Improper brakes 22 Other improper driving 23 Not applicable or not stated
68-69	Violation #2 Values same as Violation #1
70	Driver Injury Class 1 Not injured 2 Class C injury 3 Class B injury 4 Class A injury 5 Killed 6 Driver not present 7 Not stated

Position	Contents
71-74	Means of Involvement

MRSI

M - Means of Involvement

Single Vehicle Accident

1 Ran off road
(1 veh. with acc. type = 1,2,3)
2 Hit fixed object
(1 veh. with acc. type = 11)
3 Hit non-fixed object
(1 veh. with acc. type = 4,5,12)

Multi-Vehicle Accident

4 Car vs car
(2 cars of veh. type = 1,4,14,19)
5 Car vs truck or bus
(car with above veh. type &
truck of veh. type = 5 through 13)
6 More than two vehicles involved

Other Accidents

7 Any 1 or 2 veh. accident not
categorized above
(e.g., acc. type = 6,8,9,10 &
2 vehicle accidents involving
2 trucks or any motorcycles)

R - Region of Impact

1 Frontal collision
(pt. of contact = 1,2,3,4,21,25)
2 Right side collision
(p.o.c. = 18,19,20,28)
3 Left side collision
(p.o.c. = 5,6,7,26)
4 Rear end collision
(p.o.c. = 8,14,15,16,17,27)
5 Unspecified
(p.o.c. = 9 through 13 &
22,23,24,29,30,31)

S - Speed of Accident

1 00-29 mph
2 30-49 mph
3 50-79 mph
4 Not stated

Position		Contents
		I - Injury to Driver
		2 Not injured 3 Class C injury 4 Class B injury 5 Class A injury 6 Killed 7 Not stated
		2nd. Accident
75-106		Same as first accident codes. Accident year and all other variables will be 9's if accident not present.
		3rd. Accident
107-138		Same codes as for accident 1-2
139-142		Days Under Observation
143-144		Number of Good Rails
145-168		Rail Area Number 1 (Dec. 31, 1974 - July 1, 1974)
	145	Number of Speeding Convictions
	146	Number of Stop Convictions
	147	Number of Moving Convictions
	148	Number of Reckless Convictions
	149	Number of Alcohol Convictions
	150	Number of Administrative Convictions
	151	Number of Accidents at Fault
	152	Number of Suspension and Revocation
	153	Number of Equipment Convictions
	154	Number of Violation Convictions
	155	Number of Accident Violation Convictions
	156	Number of Accidents

Position		Contents
	157	Number of 4-Point Letters
	158	Number of 7-Point Letters
	159	Number of Suspensions
	160	Number of Revocations
	161	Number of Conferences
	162	Number of Hearings
	163	Number of Preliminary Hearings
	164	Number of Accidents Not At Fault
	165-167	Number of Days Under Suspension or Revocation
	168	Error Check Code (0 correct, 1-4 error)
169-336		Rail Area Numbers 2-8 (Seven 24 byte rails in 6 month periods June 30, 1974 - January 1, 1975)
337-354		1st 6 months 'Raters Rail' (Dec. 31, 1974 - July 1, 1974)
	337	Number of Violations
	338	Number of Day Violations
	339	Number of Night Violations
	340	Number of BAC's, 0 - .05
	341	Number of BAC's, .06 - .09
	342	Number of BAC's, .10 - .14
	343	Number of BAC's, .15 - .19
	344	Number of BAC's, .20 - .24
	345	Number of BAC's, .25 - .54
	346	All Other BAC's
	347	Number of Crash Involved Arrests
	348	Number of DUI's Tried

Position	Contents
349	Number of Other Offense Tried
350	Number of DUI Convictions
351	Number of Other Offense Convictions
352	Number of Not Guilty's For Noted Offense
353	Number of PJC's
354	Number of NOL PROS's
355-372	2nd. RATERS RAIL (June 31, 1974 - Jan. 1, 1974)
373	Race 1 White 2 Non-white
374	Sex 1 Male 2 Female

APPENDIX B

Predicted Probabilities of A/R Crash Involvement

Figure B-1. General population model - predicted probabilities of A/R crash involvement.

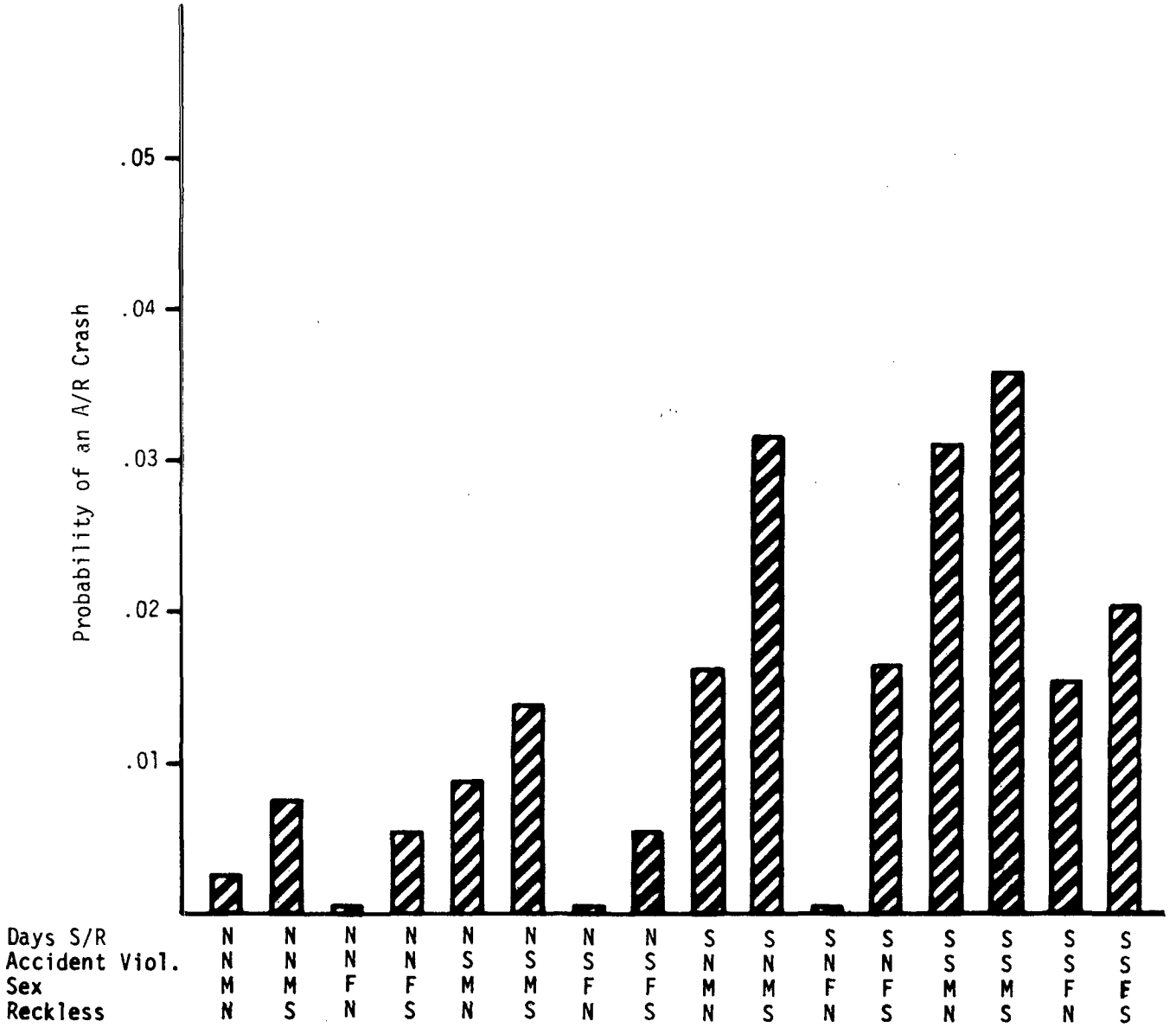


Figure B-2. Males, 16-20 model - predicted probabilities of A/R crash involvement.

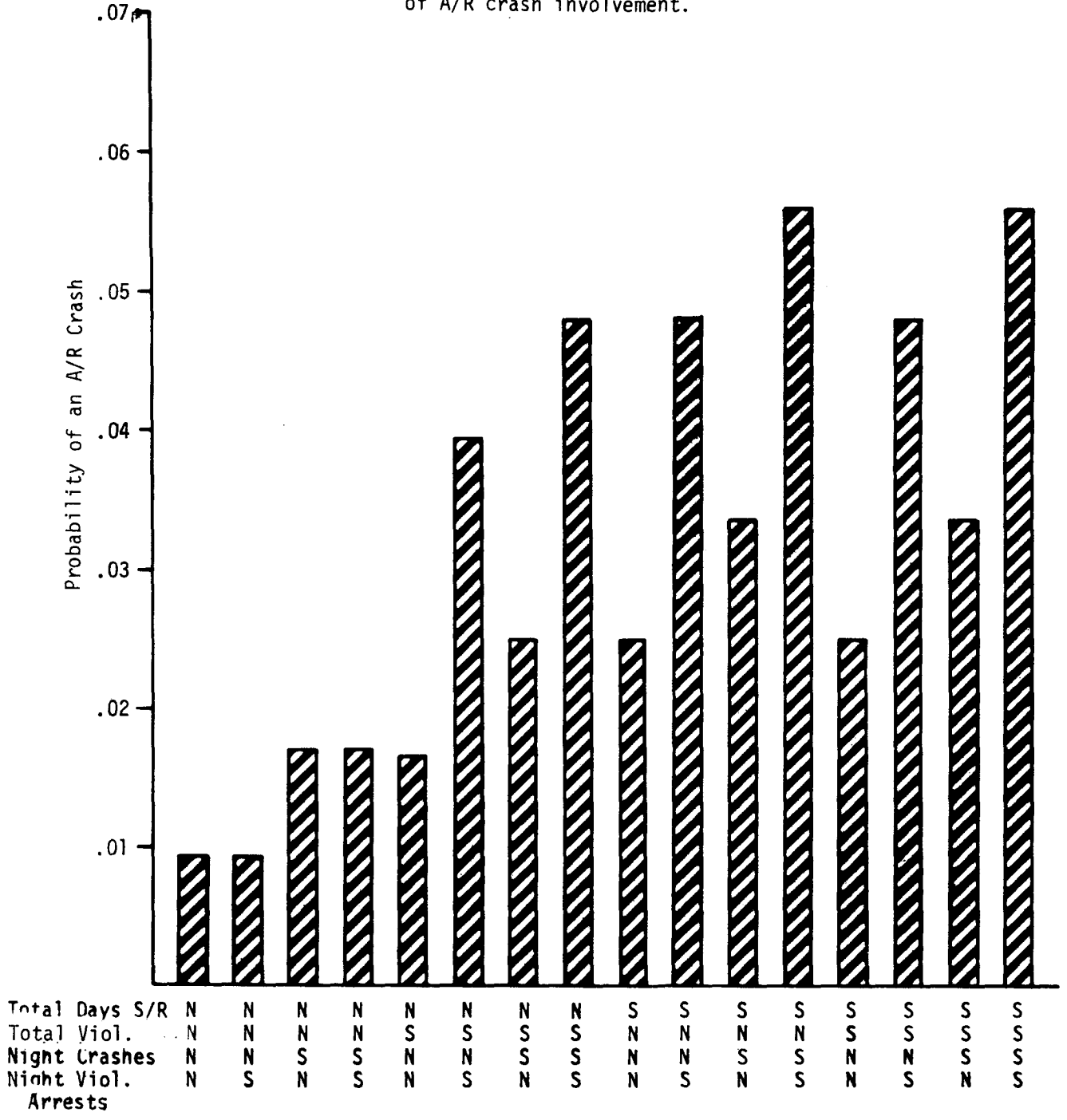


Figure B-3. Males, 21-24 model - predicted probabilities of A/R crash involvement.

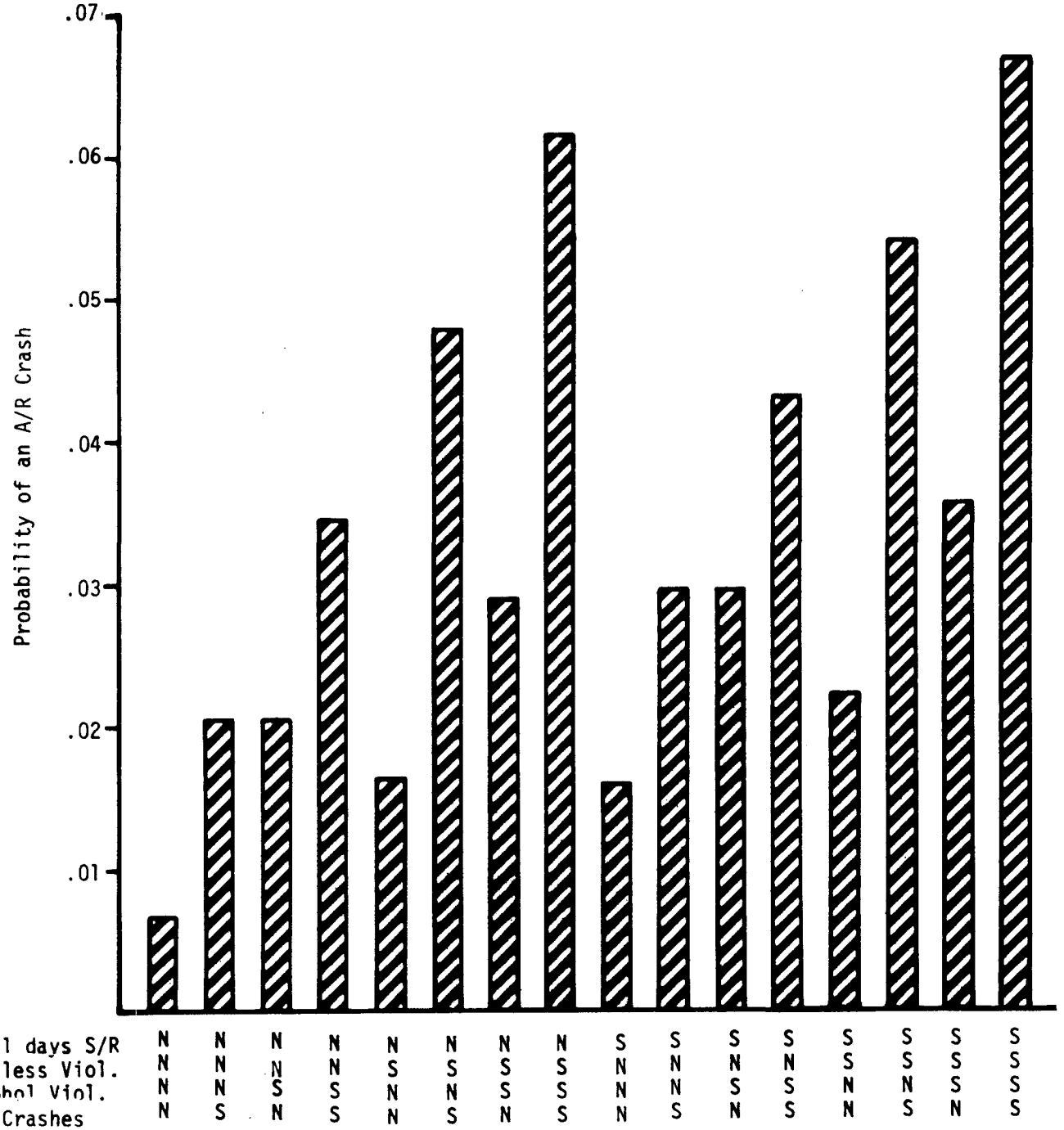


Figure B-4. DUI group model - predicted probabilities of A/R crash involvement.

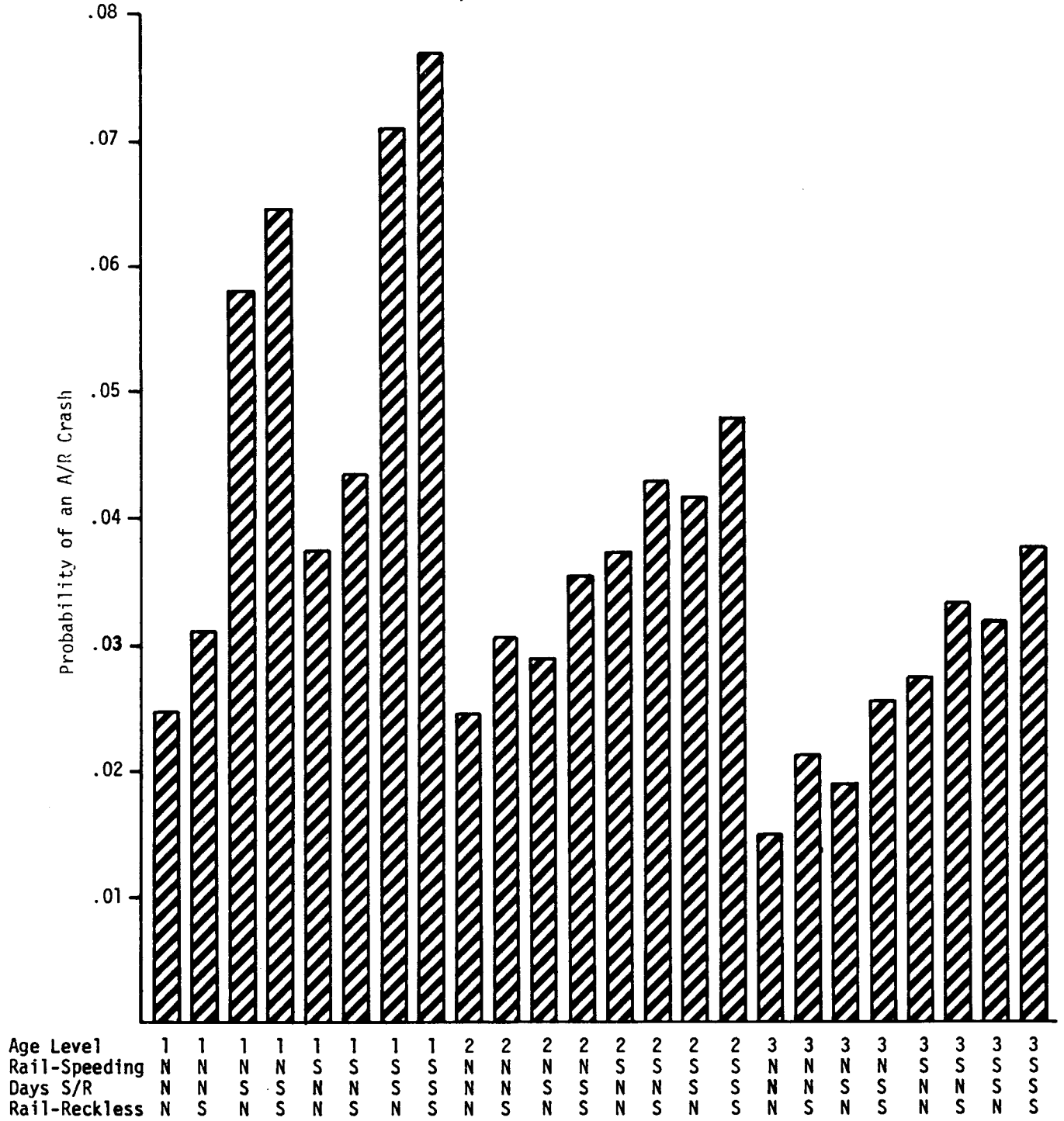


Figure B-5. Three or more violations group model - predicted probabilities of A/R crash involvement.

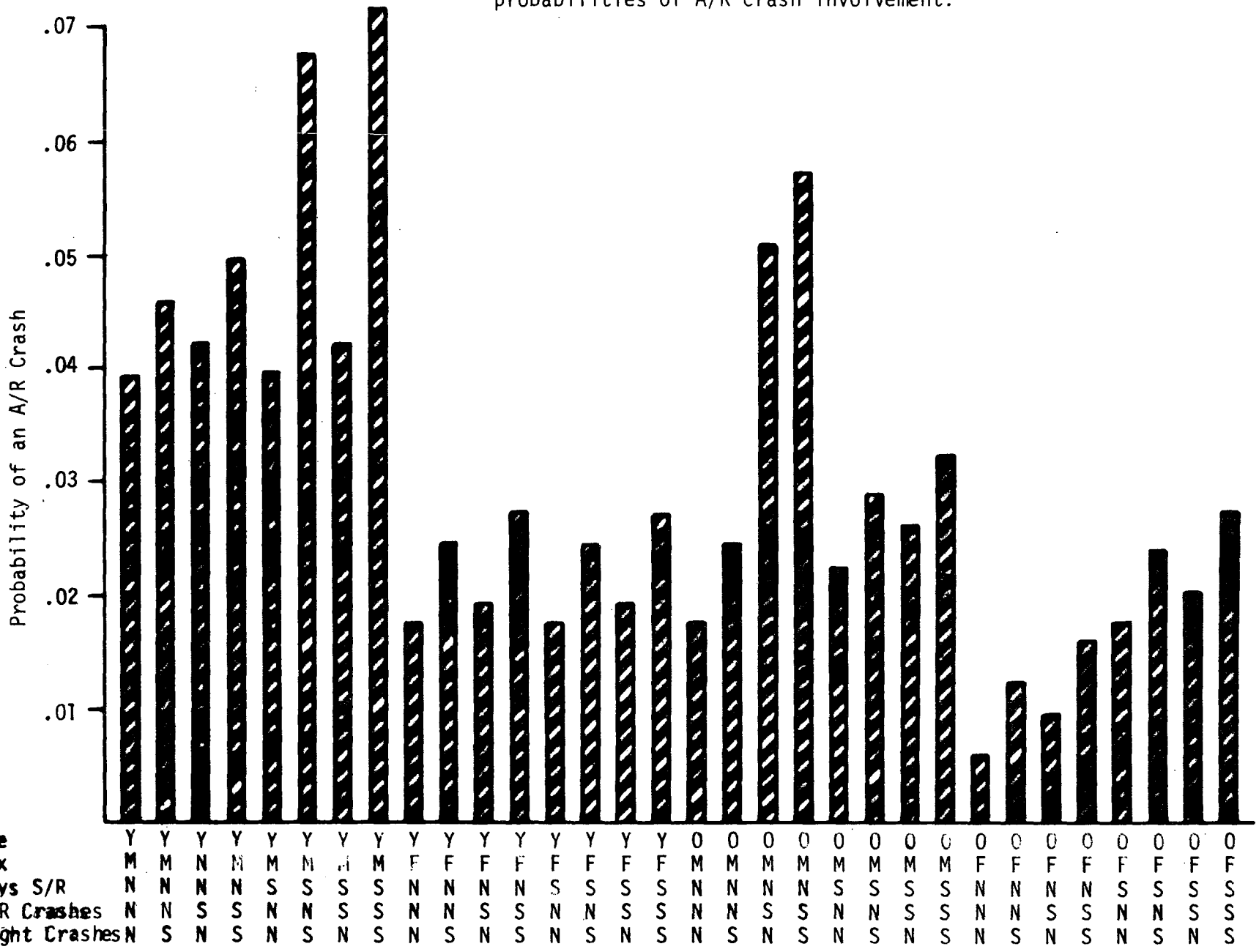


Figure B-6. Divorce group model - predicted probabilities of A/R crash involvement.

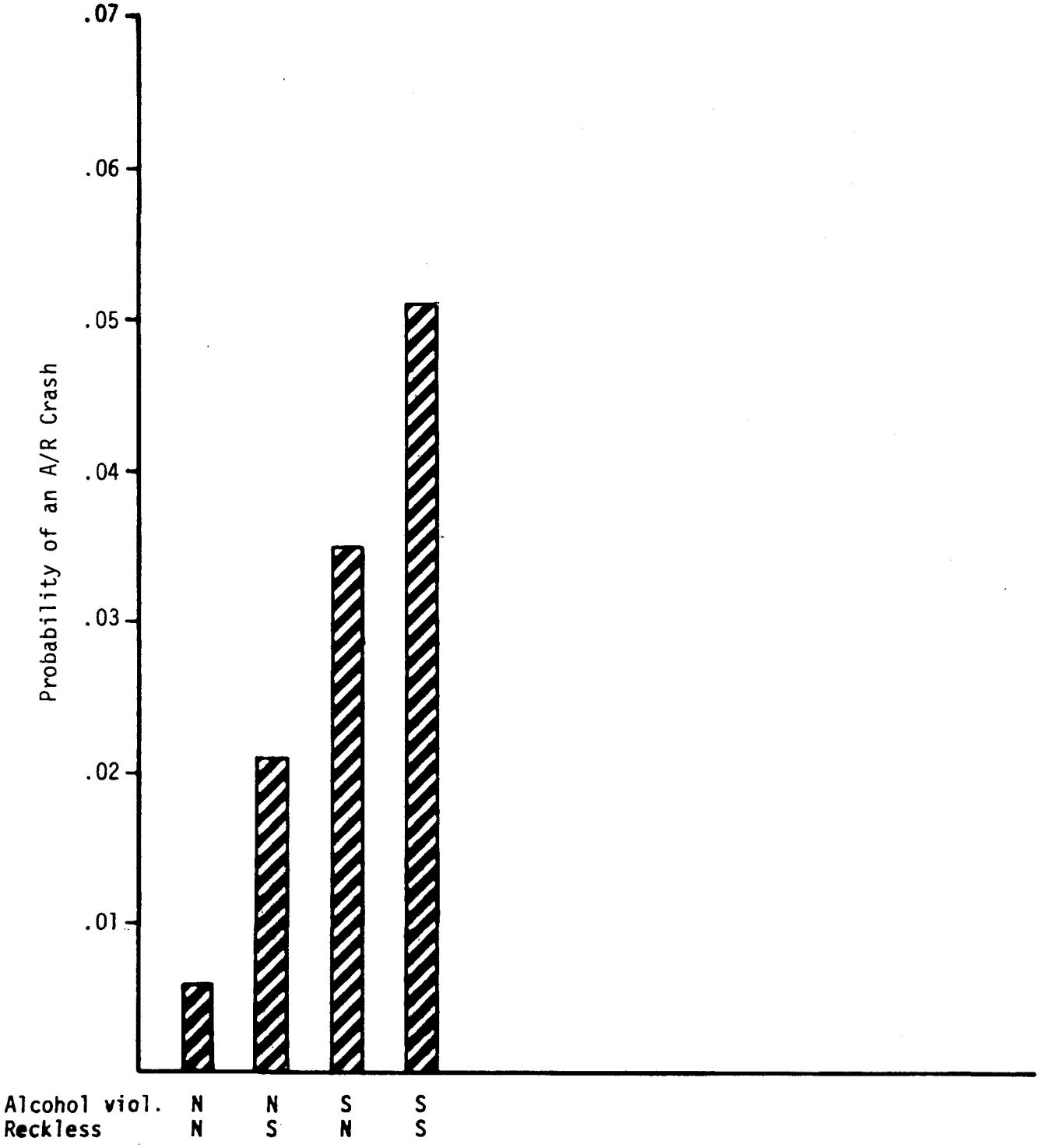
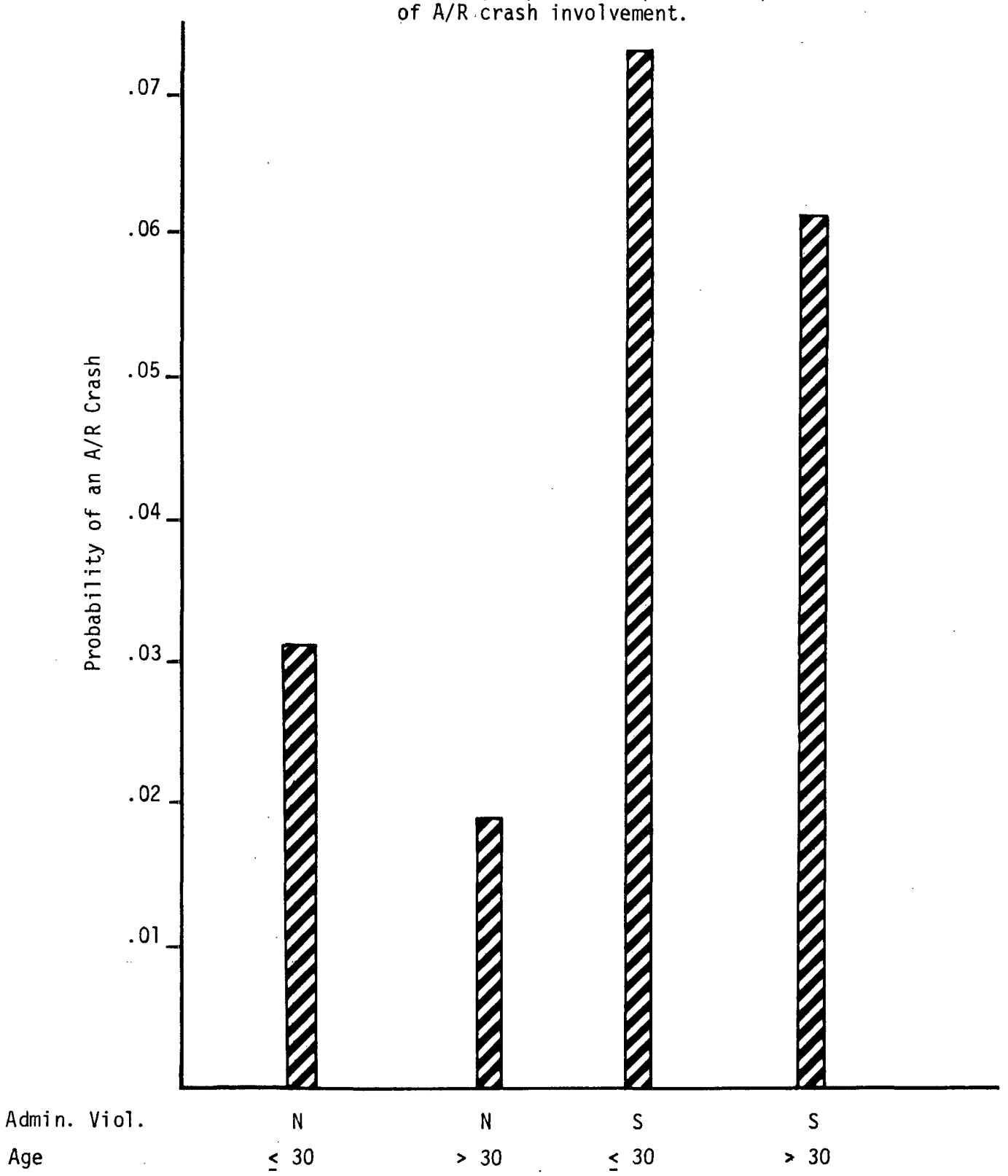


Figure B-7. Prison group model - predicted probabilities of A/R crash involvement.



APPENDIX C

Design Matrices and Model Coefficients

Figure C-1. Design matrix and model coefficients - general population model.

1	1	0	1	0	0	0	0
2	1	0	1	0	0	1	0
3	1	0	0	0	0	0	0
4	1	0	0	0	0	1	0
5	1	0	0	1	0	0	0
6	1	0	0	1	0	1	0
7	1	0	0	0	0	0	0
8	1	0	0	0	0	1	0
9	1	0	0	0	1	0	0
10	1	0	0	0	1	0	1
11	1	0	0	0	0	0	0
12	1	0	0	0	0	0	1
13	0	1	0	0	1	0	0
14	0	1	0	0	1	1	0
15	0	1	0	0	0	0	0
16	0	1	0	0	0	1	0

$$\hat{B} = \begin{bmatrix} .00050 \\ .01546 \\ .00234 \\ .00849 \\ .01565 \\ .00489 \\ .01579 \end{bmatrix}$$

χ^2 - due to model = 469.78 d.f. = 6

χ^2 - due to error = 7.63 d.f. = 9 ($p > .50$)

$R^2 = .984$

Ratio of largest predicted value to smallest = 72.0

Figure C-2. Design matrix and model coefficients - 16-20 yr. old males model.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$\hat{B} = \begin{bmatrix} .00933 \\ .01664 \\ .02533 \\ .00855 \\ .02291 \end{bmatrix}$$

χ^2 due to model = 185.40 d.f. = 4

χ^2 due to error = 10.14 d.f. = 11

($p > .50$)

$R^2 = .948$

Ratio of largest predicted value to smallest = 6.09

Figure C-3. Design matrix and model coefficients - 21-24 yr. old males model.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$\hat{B} = \begin{bmatrix} .00698 \\ .01620 \\ .02240 \\ .01353 \\ .03184 \end{bmatrix}$$

χ^2 due to model = 345.38 d.f. = 4

χ^2 due to error = 5.22 d.f. = 11

($p > .90$)

$R^2 = .985$

Ratio of largest predicted value to smallest = 9.71

Figure C-4. Design matrix and model coefficients - DUI model.

1.	1	0	0	0	0	0
2.	1	0	0	0	0	1
3.	1	0	0	1	0	0
4.	1	0	0	1	0	1
5.	1	0	1	0	0	0
6.	1	0	1	0	0	1
7.	1	0	1	1	0	0
8.	1	0	1	1	0	1
9.	1	0	0	0	0	0
10.	1	0	0	0	0	1
11.	1	0	0	0	1	0
12.	1	0	0	0	1	1
13.	1	0	1	0	0	0
14.	1	0	1	0	0	1
15.	1	0	1	0	1	0
16.	1	0	1	0	1	1
17.	0	1	0	0	0	0
18.	0	1	0	0	0	1
19.	0	1	0	0	1	0
20.	0	1	0	0	1	1
21.	0	1	1	0	0	0
22.	0	1	1	0	0	1
23.	0	1	1	0	1	0
24.	0	1	1	0	1	1

$$\hat{B} = \begin{bmatrix} .02477 \\ .01507 \\ .01261 \\ .03353 \\ .00443 \\ .00610 \end{bmatrix}$$

χ^2 due to model = 61.28

χ^2 due to error = 16.73

d.f. = 18 ($p > .50$

$R^2 = .786$

Ratio of largest predicted value to smallest = 5.11

Figure C-5. Design matrix and model coefficients - three or more violations model.

1	1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0
3	1	0	0	0	1	0	0	0
4	1	0	0	0	1	0	1	0
5	1	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	1
7	1	0	0	0	1	0	0	0
8	1	0	0	0	1	0	0	1
9	0	1	0	0	0	0	0	0
10	0	1	0	0	0	0	1	0
11	0	1	0	0	1	0	0	0
12	0	1	0	0	1	0	1	0
13	0	1	0	0	0	0	0	0
14	0	1	0	0	0	0	1	0
15	0	1	0	0	1	0	0	0
16	0	1	0	0	1	0	1	0
17	0	1	0	0	0	0	0	0
18	0	1	0	0	0	0	1	0
19	0	1	0	0	0	1	0	0
20	0	1	0	0	0	1	1	0
21	0	0	1	0	0	0	0	0
22	0	0	1	0	0	0	1	0
23	0	0	1	0	1	0	0	0
24	0	0	1	0	1	0	1	0
25	0	0	0	1	0	0	0	0
26	0	0	0	1	0	0	1	0
27	0	0	0	1	1	0	0	0
28	0	0	0	1	1	0	1	0
29	0	1	0	0	0	0	0	0
30	0	1	0	0	0	0	1	0
31	0	1	0	0	1	0	0	0
32	0	1	0	0	1	0	1	0

$$\hat{B} = \begin{bmatrix} .03946 \\ .01739 \\ .02255 \\ .00590 \\ .00342 \\ .03393 \\ .00674 \\ .02834 \end{bmatrix}$$

χ^2 due to model = 539.59 d.f. = 7

χ^2 due to error = 16.238 d.f. = 24 ($p > .75$)

$R^2 = .971$

Ratio of largest predicted value to smallest = 12.09

Figure C-6. Design matrix and model coefficients - divorce model.

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \qquad \hat{B} = \begin{bmatrix} .00570 \\ .03571 \\ .01549 \end{bmatrix}$$

χ^2 due to model = 27.53 d.f. = 2

χ^2 due to error = 0.62 d.f. = 1 ($p > .25$)

$R^2 = .978$

Ratio of largest predicted value to smallest = 8.98

Figure C-7. Design matrix and model coefficients - prison model.

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad \hat{B} = \begin{bmatrix} .0184 \\ .0602 \\ .0131 \end{bmatrix}$$

χ^2 due to model = 7.62 d.f. = 2

χ^2 due to error = 1.57 d.f. = 1 p = .21

R^2 = .829

Ratio of largest predicted value to smallest = 3.99

APPENDIX D

Prospective Validity Test Frequency Distributions

Table D-1.

General population - prospective A/R crash frequencies.

	Days S/R	Acc. Viol.	Sex	Reckless	One-Year Test Frequencies		Two-Year Test Frequencies	
					No A/R Crashes	A/R Crashes	No A/R Crashes	A/R Crashes
1.	N	N	M	N	116505	403	116254	346
2.	N	N	M	S	4451	49	4526	48
3.	N	N	F	N	113494	73	109088	48
4.	N	N	F	S	898	1	798	4
5.	N	S	M	N	6528	57	6922	65
6.	N	S	M	S	1598	31	1576	16
7.	N	S	F	N	4200	6	4162	4
8.	N	S	F	S	386	2	410	3
9.	S	N	M	N	10738	207	13144	189
10.	S	N	M	S	2272	66	2408	52
11.	S	N	F	N	995	7	1354	3
12.	S	N	F	S	88	6	93	2
13.	S	S	M	N	2244	68	2435	64
14.	S	S	M	S	1583	66	1633	68
15.	S	S	F	N	265	3	241	6
16.	S	S	F	S	71	0	82	0

Table D-2.

16-20 yr. old males - prospective A/R crash frequencies.

	Days S/R	Total Viol.	Night Crashes	Night Viol. Arrests	One-Year Test Frequencies		Two-Year Test Frequencies	
					No A/R Crashes	A/R Crashes	No A/R Crashes	A/R Crashes
1.	N	N	N	N	91509	1080	91504	1111
2.	N	N	N	S	224	15	190	11
3.	N	N	S	N	5327	110	5122	101
4.	N	N	S	S	41	1	35	2
5.	N	S	N	N	22716	487	23034	429
6.	N	S	N	S	606	34	453	17
7.	N	S	S	N	3715	121	3332	107
8.	N	S	S	S	261	12	155	10
9.	S	N	N	N	2186	89	2650	96
10.	S	N	N	S	17	0	16	0
11.	S	N	S	N	577	26	667	24
12.	S	N	S	S	9	2	18	1
13.	S	S	N	N	4744	143	5842	173
14.	S	S	N	S	857	49	771	41
15.	S	S	S	N	1409	54	1612	85
16.	S	S	S	S	401	22	267	16

Table D-3

21-24 yr. old males - prospective A/R crash frequencies.

	Total Days S/R	Reckless Viol.	Alcohol Viol.	A/R Crashes	One-Year Test Frequencies		Two-Year Test Frequencies	
					No A/R Crashes	A/R Crashes	No A/R Crashes	A/R Crashes
1.	N	N	N	N	112299	919	110095	761
2.	N	N	N	S	1122	48	1052	29
3.	N	N	S	N	385	11	387	9
4.	N	N	S	S	38	1	37	2
5.	N	S	N	N	11635	192	11636	160
6.	N	S	N	S	704	31	580	20
7.	N	S	S	N	112	4	93	5
8.	N	S	S	S	21	1	26	1
9.	S	N	N	N	13916	253	14728	227
10.	S	N	N	S	347	16	314	22
11.	S	N	S	N	4650	190	4755	150
12.	S	N	S	S	599	28	537	27
13.	S	S	N	N	7343	205	7447	167
14.	S	S	N	S	596	43	604	38
15.	S	S	S	N	1964	100	2020	87
16.	S	S	S	S	524	35	464	31

Table D-4.

DUI group - prospective A/R crash frequencies.

	Age Level	Speed. Viol.	Days S/R	Reckless Viol.	One-Year Test Frequencies		Two-Year Test Frequencies	
					No A/R Crashes	A/R Crashes	No A/R Crashes	A/R Crashes
1.	1	N	N	N	174	11	364	17
2.	1	N	N	S	33	3	52	3
3.	1	N	S	N	739	44	641	41
4.	1	N	S	S	55	3	51	3
5.	1	S	N	N	85	5	141	7
6.	1	S	N	S	18	1	26	2
7.	1	S	S	N	89	8	91	7
8.	1	S	S	S	16	1	27	2
9.	2	N	N	N	1491	57	1470	68
10.	2	N	N	S	175	6	163	8
11.	2	N	S	N	4664	193	4449	177
12.	2	N	S	S	276	17	240	12
13.	2	S	N	N	475	13	456	12
14.	2	S	N	S	62	1	76	3
15.	2	S	S	N	556	24	470	13
16.	2	S	S	S	78	4	64	2
17.	3	N	N	N	9286	206	10687	226
18.	3	N	N	S	547	17	653	17
19.	3	N	S	N	28652	618	31573	650
20.	3	N	S	S	889	32	851	35
21.	3	S	N	N	1222	31	1244	33
22.	3	S	N	S	105	4	134	2
23.	3	S	S	N	1551	59	1577	49
24.	3	S	S	S	163	9	163	10

Table D-5.

Three or more violations - prospective A/R crash frequencies.

	Age	Sex	Days S/R	A/R Crashes	Night Crashes	One-Year Test Frequencies		Two-Year Test Frequencies	
						No A/R Crashes	A/R Crashes	No A/R Crashes	A/R Crashes
1.	Y	M	N	N	N	5512	167	11226	283
2.	Y	M	N	N	S	982	40	1714	53
3.	Y	M	N	S	N	66	4	80	2
4.	Y	M	N	S	S	395	22	473	22
5.	Y	M	S	N	N	3542	144	5144	214
6.	Y	M	S	N	S	789	26	1092	51
7.	Y	M	S	S	N	121	7	117	9
8.	Y	M	S	S	S	495	34	513	34
9.	Y	F	N	N	N	569	4	1198	13
10.	Y	F	N	N	S	82	2	148	1
11.	Y	F	N	S	N	2	0	6	0
12.	Y	F	N	S	S	13	0	11	3
13.	Y	F	S	N	N	207	3	314	2
14.	Y	F	S	N	S	46	1	56	2
15.	Y	F	S	S	N	2	1	2	0
16.	Y	F	S	S	S	7	1	16	0
17.	O	M	N	N	N	82829	1321	86730	1226
18.	O	M	N	N	S	7001	181	7387	158
19.	O	M	N	S	N	1191	54	1342	49
20.	O	M	N	S	S	2987	169	3158	122
21.	O	M	S	N	N	33523	867	41599	935
22.	O	M	S	N	S	2582	84	2762	67
23.	O	M	S	S	N	1874	77	2019	76
24.	O	M	S	S	S	3433	152	3567	159
25.	O	F	N	N	N	11520	63	11105	66
26.	O	F	N	N	S	899	11	905	8
27.	O	F	N	S	N	77	3	83	2
28.	O	F	N	S	S	161	7	148	5
29.	O	F	S	N	N	1761	27	1985	14
30.	O	F	S	N	S	157	2	141	0
31.	O	F	S	S	N	99	5	102	1
32.	O	F	S	S	S	160	0	156	0

Table D-6.

Divorce group - prospective A/R crash frequencies.

	Alcohol Violations	Reckless	One-Year Test Frequencies		Two-Year Test Frequencies	
			No A/R Crashes	A/R Crashes	No A/R Crashes	A/R Crashes
1.	N	N	11410	76	10928	58
2.	N	S	722	23	763	12
3.	S	N	875	21	987	19
4.	S	S	174	8	201	5

Table D-7.

Prison group - prospective A/R crash frequencies.

	Admin. Viol.	Age	One-Year Test Frequencies		Two-Year Test Frequencies	
			No A/R Crashes	A/R Crashes	No A/R Crashes	A/R Crashes
1.	N	Y	723	20	708	17
2.	N	O	719	16	1088	22
3.	S	Y	73	5	81	3
4.	S	O	32	1	69	1