Report 03169-VI

ACQUISITION OF INFORMATION ON EXPOSURE AND ON NON-FATAL CRASHES

Volume VI-Accident Rate Analysis

Philip S. Carroll

Highway Safety Research Institute
The University of Michigan
Ann Arbor, Michigan 48104

July 31, 1971

Supplementary Final Report

Prepared under contract FH-11-7293 for

National Highway Traffic Safety Administration
Department of Transportation
Washington, D.C. 20591

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
03169-VI		

4. Title and Subtitle	5. Report Date
Acquisition of Information on Exposure and on Non-Fatal Crashes Volume VI - Accident Rate Analysis	July 31, 1971 6. Performing Organization Code
7. Author(s) P.S. Carroll	8. Performing Organization Report No.
9. Performing Organization Name and Address Highway Safety Research Institute The University of Michigan Ann Arbor, Michigan 48104	10. Work Unit No. 11. Contract or Grant No. FH-11-7293
12. Sponsoring Agency Name and Address National Highway Traffic Safety Admin. Washington, D.C. 20591	13. Type of Report and Period Covered Supplementary Final Report March, 1971 - July, 1971 14. Sponsoring Agency Code FH-11

15. Supplementary Notes

The opinions, findings, and conclusions expressed in this publication are those of the author and not necessarily those of the National Highway Traffic Safety Administration.

16. Abstract

A group accident-rate difference method is introduced as a means of identifying driver-vehicle-road-environment classes which have unique accident-rates. Following a comparison with the AID analysis method of Volume I, the new method is recommended. The new method is also applied to the data of the pilot survey described in Volume I, and modified recommendations are produced regarding variables in future exposure surveys and a hierarchy of accident-rate classes for future evaluations. The recommended variables are Driver Sex, Driver Age, Vehicle Type, Model Year, Road Type, Day/Night, and Vehicle Make, where addition of the latter variable is the only change in the previous recommendations. The new hierarchy contains only 18 driver-vehicle-road-environment classes, compared to 26 classes in the previous hierarchy.

Three methods are compared for collection of accident data to be used in accident-rate calculation (mass accident data from state records, individual state records of exposure survey subjects, and accident involvements recalled by survey subjects). The first method is recommended because of greater accuracy.

After consideration of methods for combining independent sources of exposure data, it is recommended that data from a single, comprehensive survey be used in calculating group accident rate.

17. Key Words	18. Distribution Statement
	NTIS For public distribution

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
none	none	55	

Form DOT F 1700.7 (8-69)

Preface

This is a supplementary report on Contract FH-11-7293, "Acquisition of Information on Exposure and on Non-Fatal Crashes," covering work on accident rate analysis beyond Volumes I-V of the final report.

Requirements, approaches, and findings are presented for each of the supplementary tasks. Final conclusions and recommendations are presented with regard to classes of driver-vehicle-road-environment combinations for which accident data and exposure data should be obtained in order to determine accident rates for national, highway safety evaluations.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION 1
INTRODUCTION

In Volume I of the final report on this contract, unique classes of driver-vehicle-road-environment combinations were identified on the basis of maximum homogeneity of the distributions of exposure values in those classes. The classes were recommended as ones for which accident rates should be determined in the future for national evaluations of highway safety countermeasures. The analysis procedures for selecting the unique classes were based on exposure data obtained for each subject in the pilot survey described in Volume I. Because the exposure data (estimate of vehicle miles travelled in the past 30 days) was the primary dependent variable in the survey, it was the logical basis for data classification.

However, the pilot survey also included the collection of data on accident involvements recalled by the subjects during the preceeding 3 years. Furthermore, the data file includes a computed accident rate for each subject, derived as the quotient of recalled accident frequency and estimated mileage. Over 20 percent of the subject cases have a non-zero value of accident rate, and the remainder have zero values. Thus, the opportunity exists for the analysis of unique classes with accident rate as the dependent variable, rather than exposure. This is a more logical approach, because the uniqueness of accident rates among classes is the ultimate objective, rather than the uniqueness of exposure. Previously, it had been assumed that average exposure and average accident rate would be approximately proportional among classes, and that either dependent variable could be used for the classification analysis.

Two problems with the data led to an earlier rejection of accident rate as the dependent variable. First, there was an

apparent underreporting of accidents by survey subjects, probably due to forgetting or reluctance to admit accidents in an interview. The biases inherent in the underreporting are unknown. Second, the time periods of the exposure and accident estimates do not coincide (past 30 days vs. past 3 years). One recourse is to assume uniform exposure for each subject over a 3 year period, and to extrapolate the individual 30-day estimates to 3 years (a factor of 36 months). An individual accident rate may then be calculated for each subject for the "equivalent 3-year period." Some cases will have misleading individual accident rates because of atypical mileage in the past 30 days. Nevertheless, a mean value of individual accident rates may be calculated for the whole data sample resulting in 35 accidents per million miles. On the other hand, a group accident rate may be calculated for the whole sample by dividing the total number of admitted accidents by the total number of miles (extrapolated to 3 years). The resulting group accident rate is only 8.4 accidents per million miles--one quarter the mean value of individual accident rates. The discrepancy is due to the unusual distribution of individual accident rates, i.e. a relatively low proportion of high individual accident rates and a high proportion of zero accident rates. Over a longer period than 3 years, the same sample should have a higher proportion of non-zero individual accident rates and a lower proportion of zero accident rates. In previous studies of accident data over longer periods (e.g. 6 years) it has been shown that mean values of individual accident rates tend to coincide with group accident rates. However, because the pilot survey data covers only 3 years, its mean values of accident rates are very misleading.

There is no solution to the problem of underreporting bias in the accident data of the survey. However, there is a solution to the problem of misleading mean values of accident rate. The solution is to use an analysis methodology for selecting data classes which does not rely on mean values of accident rate, as

in the AID computer program of Volume I. The alternative is to divide data groups into subgroups which have a maximum relative difference between their group accident rates.

On this basis, the supplementary study was initiated, with the primary objective of verifying the data classes and survey variables recommended in Volume I. The supplementary work statement is presented in Appendix N.

In Section 2, the alternative methodologies for selecting accident-rate classes are compared. In Section 3, variables for exposure surveys are discussed. Section 4 considers alternate sources of data for accident-rate computations.

SECTION 2
DETERMINATION OF UNIQUE ACCIDENT-RATE CLASSES

This section covers Tasks 1-6 of the supplementary work statement. Its basic purpose is to compare alternative methodologies for deriving unique accident-rate classes and to recommend the preferred methodology.

METHODOLOGIES FOR DERIVING ACCIDENT-RATE CLASSES

The "AID analysis" method and the "group accident-rate difference" method are to be compared. Both depend on the existing data file from the exposure survey, as described in Volume I. The file was condensed to 5720 subject cases which have responses to both the mileage estimate for the past 30 days and the accident frequency for the past three years. There were 91 cases with zero miles, and therefore only 5629 cases had accident rate values. Each case included 12 independent variables (Table 1); these variables were selected from the 21 best predictors of the previous study, excluding the 9 variables which do not correspond to variables on standard accident reports.

AID Analysis

The AID analysis method is described in Volume I, as applied to exposure data. In this section, the AID methodology will be applied to accident-rate data as the dependent variable, and it will identify unique classes of the independent variables. The total sample will be divided first into the two groups (of one of the independent variables) which produce the minimum unexplained variability bewteen their respective distributions of accident-rate values. Each of the resulting groups will be further split in the same way until the divisions are no longer significant.

Table 1

Independent Variables

Driver Age
Driver Sex
Vehicle Type
Passenger Car Size
Passenger Car Manufacturer
Vehicle Model Year
Percent Driving on City Streets
Percent Driving on Urban Freeways
Percent Driving on Rural Freeways
Percent Driving on Rural Roads
Percent Driving at Night
Percent Driving on Wet Pavement

The criterion of minimum unexplained variability tends to produce divisions with maximum relative difference between the mean values of the dependent variable for the two resulting groups.

Group Accident-Rate Difference

The group accident-rate difference method is a step-by-step process of finding driver-vehicle-road-environment classes which are the best predictors of group accident rate. The first step is to determine the total number of accidents and total mileage for each level of each independent variable in the whole data set. For each variable, the levels are grouped in all possible two-group combinations which have logical meanings (e.g. old drivers and young drivers would not be grouped together). The group accident rates are determined for each group in the various combinations (total accidents for the group divided by total mileage for the group). The relative difference in group accident rate is determined for each combination. The relative difference (actual difference between the two groups divided by their average) is a better indicator of uniqueness than actual difference. For each variable, one two-group combination will have the highest relative difference. We pick a combination only if its smallest group is at least 10 percent of the population. For all 12 variables, one of them will have a combination with a relative difference higher than all the other combinations in the other variables. We then select that variable as the desired "first splitting variable" for the total sample. Thus, the sample is divided into the two groups with highest relative difference in accident rate.

The second step is to repeat the same process as above for each of the two new groups. Again we pick a combination only if it is at least 10 percent of the group above it. The result of the second step is two new second-level splitting variables and hence four new groups. It is conceivable that the two new

splitting variables would actually be the same variable, even though they would be splitting two different groups from above.

The third step again repeats the process above for each of the four new groups. In succeeding steps, the number of groups could theoretically increase to eight, sixteen, etc. However, the process is stopped whenever a group size is less than ten percent of the total sample. The smallest groups at the end of the process are unique classes of driver-vehicle-road-environment combinations which have the maximum relative homogeneity with respect to group accident rate.

In each step, the mean mileage and number of accidents for each variable level are determined by computer, and then group accident rates and relative differences are found by desk calculator. Each step begins with a filtering of the data file for the variable levels which define the newly selected group.

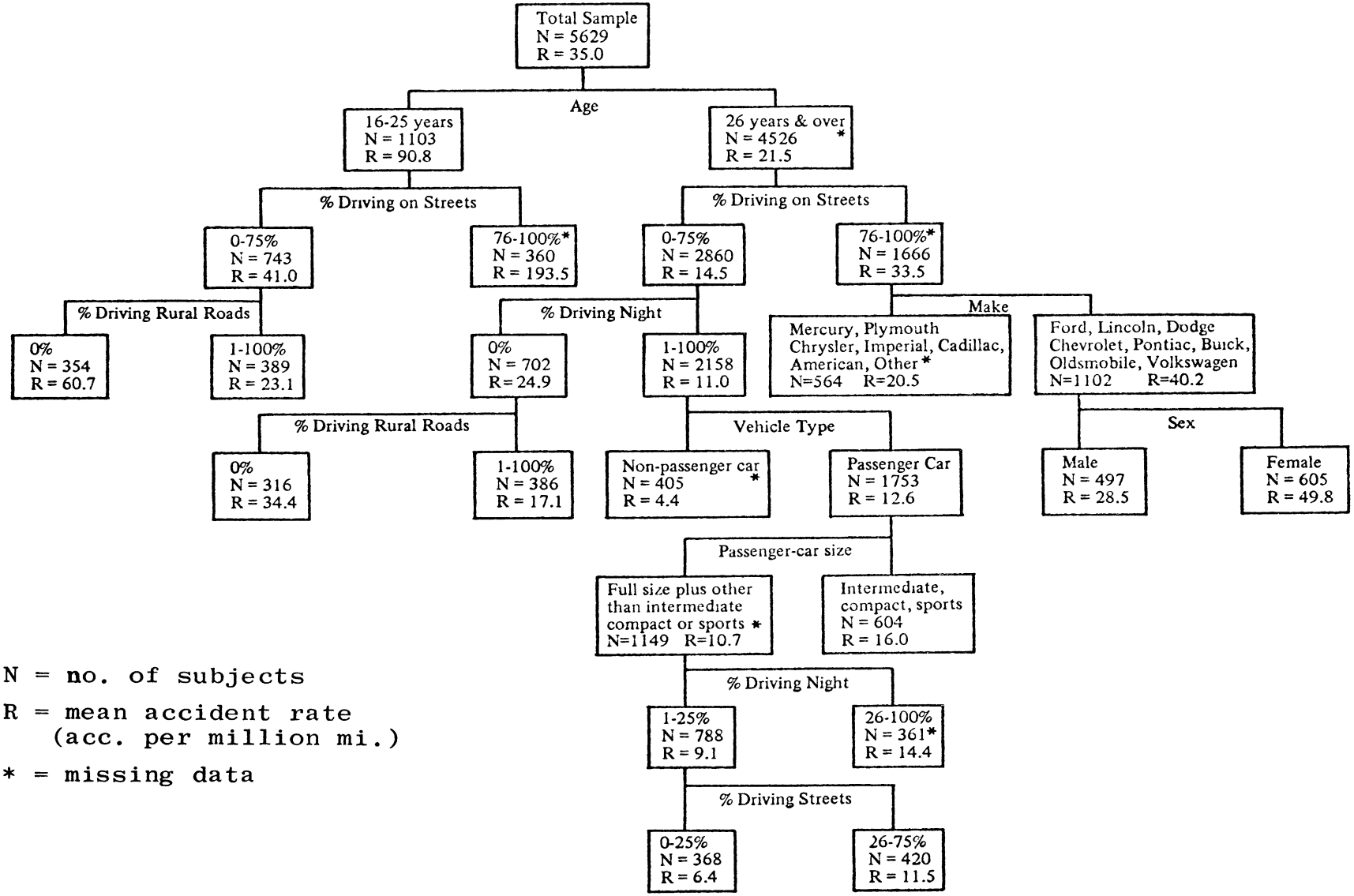
ALTERNATIVE HIERARCHIES OF ACCIDENT-RATE CLASSES

The hierarchy of accident-rate classes based on the "AID analysis" method is shown in Figure 1 and the hierarchy of accident-rate classes based on the "group accident-rate difference" method is shown in Figure 2.

Figure 1 involves 8 independent variables at 7 splitting levels with 11 intermediate and 13 final data classes. Figure 2 involves 6 independent variables at 7 splitting levels with 12 intermediate and 15 final data classes. The five common variables of the two hierarchies are as follows:

Driver Age
Percent Driving on City Streets
Driver Sex
Vehicle Type
Vehicle Make

It is clear that these five variables should be seriously considered for a recommended hierarchy of accident-rate classes. Figure 1



N = no. of subjects

R = mean accident rate
(acc. per million mi.)

* = missing data

Figure 1 -- Accident-Rate Classes Based on AID Analysis

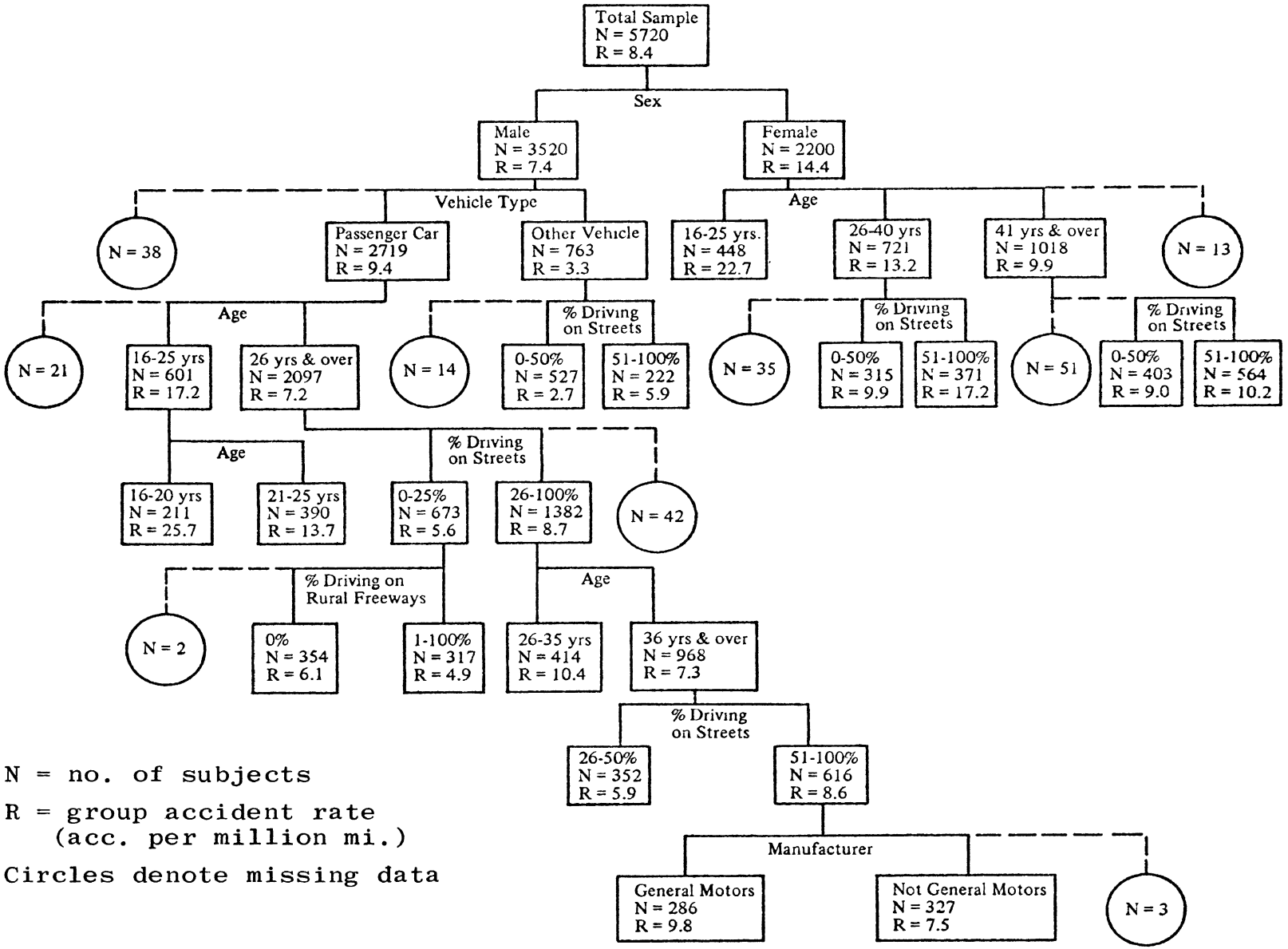


Figure 2 -- Accident-Rate Classes Based on Group Accident-Rate Differences

also involves Percent Driving on Rural Roads, Percent Driving at Night, and Passenger Car Size, while Figure 2 also involves Percent Driving on Rural Freeways.

The two hierarchies are quite different structurally in terms of the order in which the independent variables define group splits. However, Driver Age and Percent Driving on City Streets are highly important in both hierarchies. Both are unsymmetrical in that only one branch in each goes to the seventh level. Both have four final classes that are defined by five variables, and both have just one final class defined by only two variables. However, Figure 2 has eight final classes defined by three variables. None of the final classes are similar to one in the other hierarchy.

As noted previously, the major difficulty with the hierarchy produced by the "AID analysis" method is that the mean values of individual accident rates in the various groups are much larger than the corresponding group accident rates. Comparisons of these differences are shown in Tables 2 and 3. For each of the groups in Figure 1, a group accident rate was computed independently of the "AID analysis" method, and for each of the groups in Figure 2, a mean accident rate was computed independently of the "group accident-rate difference" method. In Table 2, the ratios of mean accident rate to group accident rate range from 1.6 to 8.1, while in Table 3, the ratios range from 1.6 to 5.6. In both cases, the ratios at the lower levels of the hierarchies tend to be smaller than at the higher levels. But neither hierarchy shows a clear trend toward even approximate equivalence between the means of accident rate and group accident rates.

For each of the hierarchies, statistical significance levels were determined for each of the two-way splits. The significance levels were determined from F values in the AID analysis for Figure 1 and from T values in analysis-of-variance runs for Figure 2. The significance levels are shown in Tables 2 and 3

Table 2

Groups of AID Analysis Hierarchy

Chart Level	Age	% Drive Street	% Drive Rur.Rd.	% Drive Night	Make	Veh. Type	Sex	Car Size	Mean	Grp. Ave.	Ratio	Signif.
1									35.0	8.3	4.2	1.000
2	16-25								90.8	16.2	5.6	.999
2	26+ *								21.5	6.5	3.3	.999
3	16-25	0-75							41.0	14.4	2.8	.998
3#	16-25	76-100*							193.5	23.8	8.1	-
3	26+ *	0-75							14.5	5.6	2.6	1.000
3	26+ *	76-100*							33.5	10.1	3.3	.975
4#	16-25	0-75	0						60.7	17.8	3.4	-
4#	16-25	0-75	1-100*						23.1	12.3	1.9	-
4	26+ *	0-75		0					24.9	6.4	3.9	.978
4	26+ *	0-75		1-100*					11.0	5.5	2.0	.999
4#	26+ *	76-100*			I				20.5	7.7	2.7	-
4	26+ *	76-100*			II				40.2	11.8	3.4	.939
5#	26+ *	0-75	0	0					34.4	7.1	4.5	-
5#	26+ *	0-75	1-100	0					17.1	5.5	3.1	-
5#	26+ *	0-75		1-100*		Not Car			4.4	2.7	1.6	-
5	26+ *	0-75		1-100*		Car			12.6	6.9	1.8	.997
5#	26+ *	76-100*			II		M		28.5	9.7	2.9	-
5#	26+ *	76-100*			II		F		49.8	27.7	1.8	-
6	26+ *	0-75		1-100*		Car		Full, Other*	10.7	6.1	1.8	.994
6#	26+ *	0-75		1-100*		Car		Int. Comp. Sport	16.0	8.7	1.8	-
7	26+ *	0-75		1-25		Car		Full Other *	9.1	5.4	1.7	.997
7#	26+ *	0-75		26-100*		Car		Full Other *	14.4	7.5	1.9	-
8#	26+ *	0-25		1-75		Car		Full Other *	6.4	4.1	1.6	-
8#	26+ *	26-75		1-25		Car		Full Other *	11.5	7.1	1.6	-

- Final Class; * - missing data
I - Mercury, Plymouth, Chrysler, Imperial, Cadillac, American, Other, Missing
II - Ford, Lincoln, Dodge, Chevrolet, Pontiac, Buick, Oldsmobile, Volkswagen

Table 3

Groups of Group Accident-Rate Difference Hierarchy

Chart Level	Sex	Veh. Type	Age	% Drive Street	% Drive Rur.Fwy.	Make	Grp. Ave.	Mean	Ratio	Signif.
1							8.4	35.0	4.2	.999
2	M						7.4	25.2	3.4	.973
2	F						14.4	51.0	3.5	1.000
3	M	Car					9.4	29.4	3.1	.999
3	M	Not Car					3.3	6.5	2.0	.963
3#	F		16-25				22.7	126.9	5.6	-
3	F		26-40				13.2	35.9	2.7	.903
3	F		41+				9.9	28.7	2.9	.583
4	M	Car	16-25				17.2	73.0	4.2	.947
4	M	Car	26+				7.2	16.9	2.3	.993
4#	M	Not Car		0-50			2.7	5.4	2.0	-
4#	M	Not Car		51-100			5.9	9.5	1.6	-
4#	F		26-40	0-50			9.9	23.8	2.4	-
4#	F		26-40	51-100			17.2	43.8	2.5	-
4#	F		41+	0-50			9.0	20.3	2.3	-
4#	F		41+	51-100			10.2	26.3	2.6	-
5#	M	Car	16-20				25.7	136.2	5.3	-
5#	M	Car	21-25				13.7	38.4	2.8	-
5	M	Car	26+	0-25			5.6	10.1	1.8	.458
5	M	Car	26+	26-100			8.7	19.7	2.3	.881
6#	M	Car	26+	0-25	0		6.1	11.1	1.8	-
6#	M	Car	26+	0-25	1-100		4.9	9.1	1.9	-
6#	M	Car	26-35	26-100			10.4	25.4	2.4	-
6	M	Car	36+	26-100			7.3	17.2	2.4	.902
7#	M	Car	36+	26-50			5.9	12.0	2.0	-
7	M	Car	36+	51-100			8.6	20.2	2.3	.954
8#	M	Car	36+	51-100		GM	9.8	27.9	2.8	-
8#	M	Car	36+	51-100		not GM	7.5	13.5	1.8	-

- Final Class

opposite the groups which produced the two-way splits. The significance levels of the AID analysis method tend to be higher, with only one out of 12 being less than .95. In the group accident-rate difference method, five out of 13 splits have significance levels less than .95.

PREFERRED METHODOLOGY

Both of the alternative methodologies presented above for the determination of unique accident-rate classes result in hierarchies that are very acceptable intuitively. That is, they both involve variables that are commonly used in accident research, they both involve logical groupings of variable levels, and they both indicate natural interactions among the variables.

The classes derived by the methodologies are required to be relatively homogeneous with respect to relevant exposure factors, amenable to sampling procedures, and useful for studying the impact of safety countermeasures. Both of the alternative methodologies lead to hierarchies of classes which appear to satisfy these objectives. In order to select an alternative, it will be necessary to determine that one methodology satisfied the objectives (or other criteria) better than the other.

Since both hierarchies involve nearly the same variables, they are equally amenable to sampling procedures. But they differ in regard to relative homogeneity of classes and usefulness in countermeasure evaluation.

In regard to relative homogeneity of the classes produced by the two methodologies, the AID method is slightly better, theoretically, as seen by the higher significance levels of Table 2, in comparison to Table 3. The reason is that the AID method maximizes the explained variance between groups, rather than only maximizing the difference in mean value of the dependent variable between groups. However, the data biases due to the low percentage of accidents reported in the survey probably affect the homogeneity

of the classes in the two hierarchies in different ways. Thus, the slight theoretical superiority of the AID analysis method in creating homogeneous classes may be neutralized. At this time, there is no way to determine the effects of accident-underreporting bias on the homogeneity of classes.

In regard to usefulness in countermeasure evaluation, the group accident-rate difference method is preferrable because it is based precisely on the same measure as would be used in evaluations, i.e. group accident rate. On this basis the group accident-rate difference method appears to be the best methodology for determining accident-rate classes from the available data file.

It should be noted that the exposure data used in creating the two alternative accident-rate hierarchies was obtained by means of driver's estimates of gross mileage over the past 30 day period. These estimates were not thoroughly classified by road-environment combinations within each driver's estimate, but rather, each driver gave independent estimates of percentages of his driving done under certain road and environment conditions. Thus, in the analyses leading to the accident-rate hierarchies, the classifications were made on the basis of drivers as the entities in certain "percent driving" categories, instead of being on the basis of driver-vehicle-road-environment combinations as the entities. However, when new data is obtained from trip-log exposure surveys, the latter case will be true. At that time, the accident-rate hierarchy may change because of the change in data format.

If new trip-log exposure data is accompanied by improved accident data, the hierarchy may tend to change simultaneously due to changes in both numerator and denominator of accident-rate computations. Improvements in the accident data accompanying exposure data may come about due to increased sample size, better techniques for getting subjects to admit past accidents, and longer time periods of accident recall, e.g. past 6 years instead of only past 3 years.

Regardless of the methodology chosen, it will be desirable to make some adjustments in the resulting hierarchy. Generally, the adjustments will produce symmetry and completeness in the hierarchy without changing any of the unique classes. Symmetry may often be achieved by interchanging the levels of two variables on one branch; this may cause one of the variables to be at the same level as it already was on the parallel branch. When the hierarchy is relatively symmetric, comparative evaluations among related groups will be more readily performed. The completeness of a hierarchy may be improved by adding one or two splitting variables near the bottom of certain branches. (The added variables would be ones which already appeared on one or more other branches.) Also, the splitting levels of certain variables may be changed slightly to make them consistent on adjacent branches without destroying the uniqueness of the classes.

Based on its usefulness in countermeasures evaluations and its more accurate accident-rate values, the "group accident-rate difference" method is preferred at the present time. In the future, the "AID analysis" may be preferred for annual re-assesments of accident-rate classes since is is quicker to do, but it would be dependent on continued improvements in accident experience responses of subjects in annual exposure surveys.

SECTION 3
VARIABLES FOR EXPOSURE SURVEYS

In the hierarchy of Figure 2, produced by the preferred "group accident-rate difference" method, the six independent variables involved should be recommended for inclusion in exposure surveys if feasible and meaningful. Three of the variables - Sex, Vehicle Type, and Age - are perfectly suitable. However, the variables Percent Driving on Streets and Percent Driving on Rural Freeways could not appear as such in an exposure survey based on trip logs. Instead, each trip should be classified by Road Type (streets vs. other roads). Also, there is some question about the meaning of the variable Vehicle Make (Manufacturer) because it appears at the very end of only one branch on the hierarchy.

In Figure 3, an AID hierarchy of exposure classes is presented, based on the condensed data file, with "miles driven in the past 30 days" as the dependent variable. This chart is similar to ones produced in the previous study, as presented in Volume I and Volume IV (Appendix E). Five of the six variables in Figure 3 are the same ones recommended in Volume I (Sex, Vehicle Type, Percent Driving on Streets, Model Year, Age). However, Figure 3 also includes Vehicle Make (which was not previously recommended) but it excludes Percent Driving at Night (which was recommended). The groupings of Vehicle Make in Figure 3 occur on only one branch and they do not appear to have any logical relationship, nor do they coincide with the Vehicle Make groupings of Figure 2.

In Figure 4, an AID hierarchy of accident classes is presented, again based on the condensed data file, with "accidents in the past 3 years" as dependent variable. This chart does not involve Vehicle Type or Percent Driving at Night, but it does include Vehicle Make and Model Year. Again, the Vehicle Make groups (on three branches) have no logical relationships.

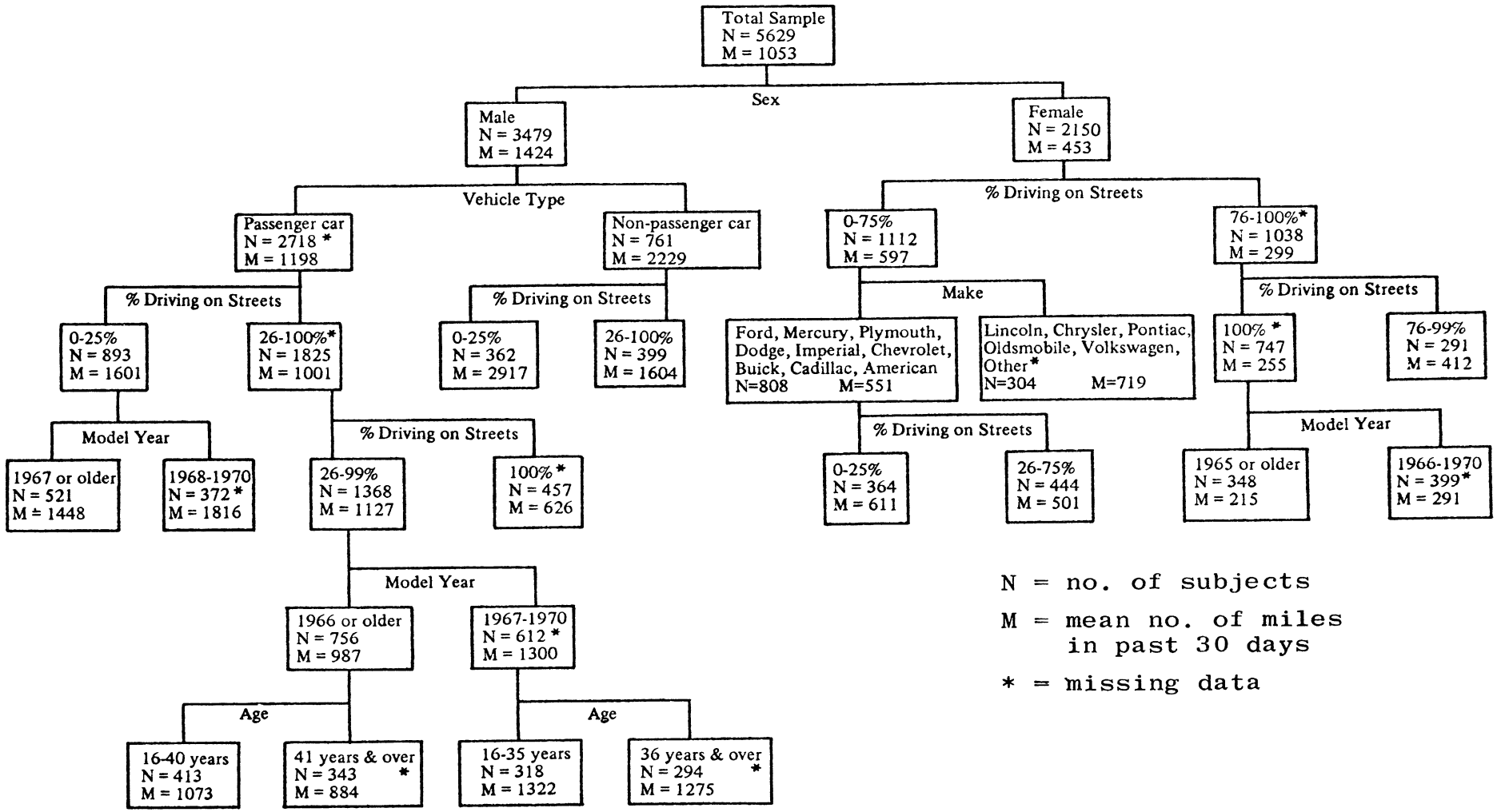


Figure 3 -- Exposure Classes Based on AID Analysis

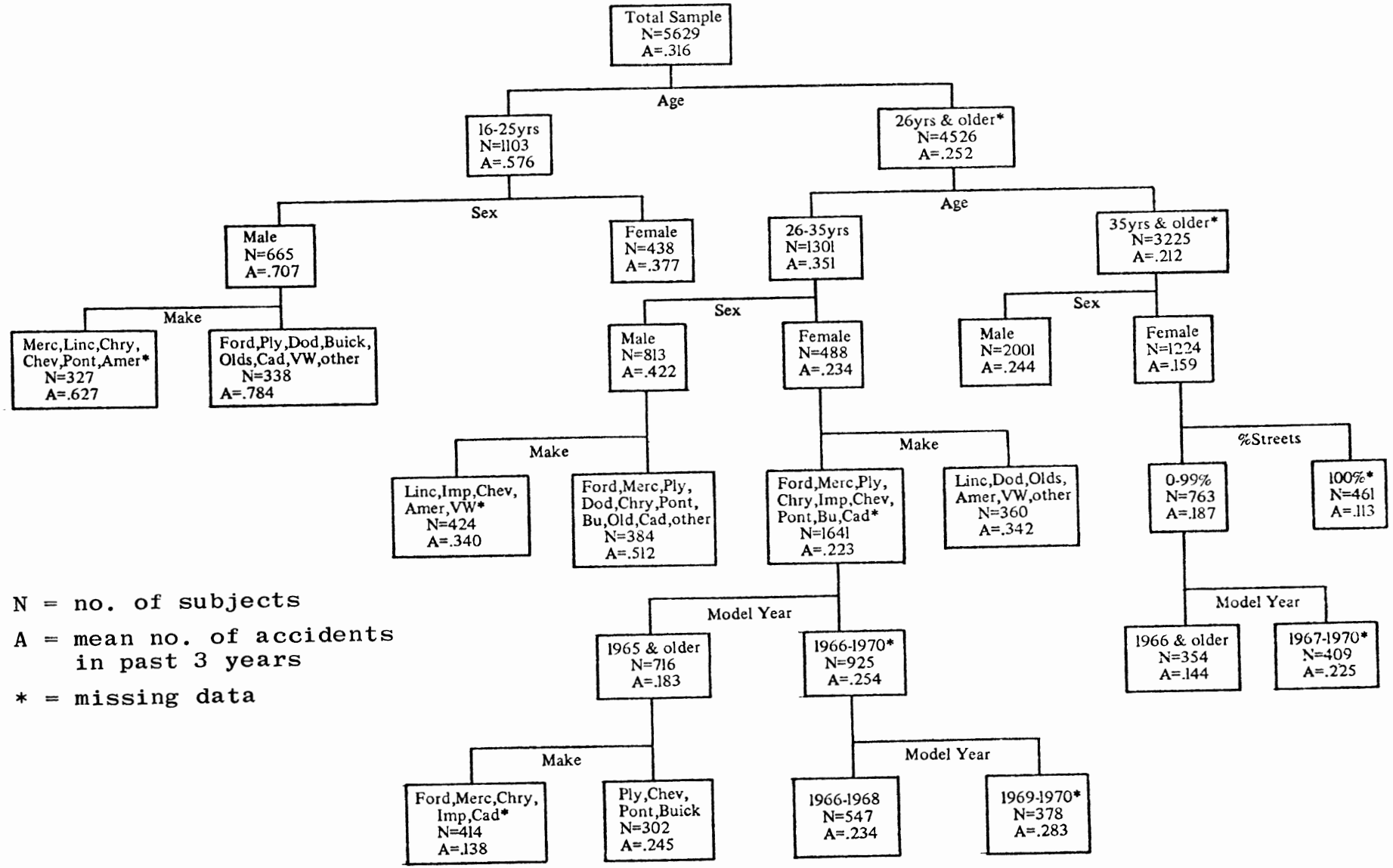


Figure 4 -- Accident Classes Based on AID Analysis

Because of the uncertainties as to proper groupings of Vehicle Make, it is concluded that a hierarchy of accident-rate classes for evaluation purposes should not include Vehicle Make. However, Vehicle Make should be included as a variable in exposure surveys for continued analysis and possible future use in classifications.

Also, it is concluded that Model Year and Day/Night should be included in the hierarchy because of strong indicators of their importance in several of the hierarchies considered.

The list of recommended variables for future exposure surveys is as follows:

Driver Sex
Driver Age
Road Type
Day/Night
Vehicle Type
Model Year
Vehicle Make

All except Vehicle Make should be included in the hierarchy of accident-rate classes for evaluations in the first year of data collection.

The recommended hierarchy of accident-rate classes is shown in Figure 5. It corresponds to Figure 2 through the third level. Slight modifications are made below that level for symmetry with respect to the Age and Road Type variables. The Day/Night and Model Year variables are added at the bottom on branches which have larger groups, and which coincidentally also show interaction with these two variables in the previous hierarchies.

The recommended hierarchy of Figure 5 includes 18 final classes of driver-vehicle-road-environment combinations, compared to the 26 classes recommended in Volume I. The main reason for the reduction in classes is the fact that the Day/Night variable now occurs on only two branches at the lowest splitting level, instead

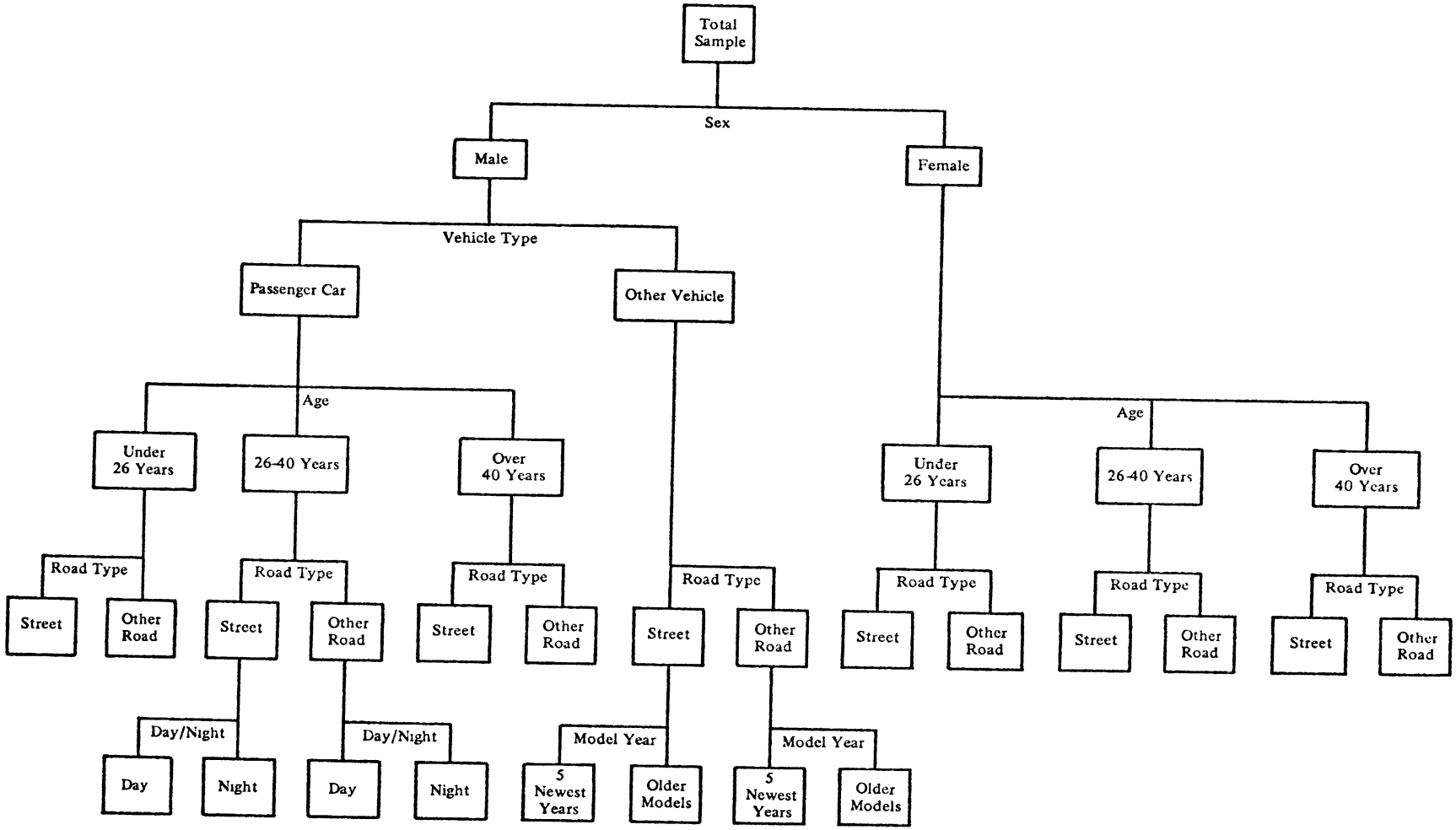


Figure 5 -- Recommended Accident-Rate Hierarchy

of four branches at the third level. This change will help to make the classes more uniform in size. Another change with the same result is use of 40 years as the dividing point in age between the middle age groups and the older age groups. The previous selection of 60 years as the dividing point was arbitrary, whereas the new dividing point of 40 years is consistent with several of the more recent hierarchies.

SECTION 4
ALTERNATIVE SOURCES OF ACCIDENT-RATE DATA

In the exposure study of Volume I, the premise was that the accident experience of many different classes of drivers and vehicles is available from current sources, but the corresponding exposure measurements for these classes are not available. Thus, the main objective was to develop procedures for collection of exposure data so that it could be combined with available accident data to provide accident rates. Although most current sources of accident data are unreliable, the accident records in official state files are widely used for accident analysis, and a natural assumption is that these statewide mass accident data would be combined with new exposure data. Likewise, a natural assumption with respect to new exposure data is that it would be collected in a single, comprehensive exposure survey. These two assumptions will be considered in the following sections.

ALTERNATIVE SOURCES OF ACCIDENT-INVOLVEMENT DATA

In Volume II, the reliability of accident involvement data was studied by means of comparisons between two different sources of accident data. One source was the incidental data obtained from subjects in the primary exposure survey, in which they gave estimates of the number of accidents in which they were involved during the past three years. The second source of data was a search of official state accident records of a subset of the survey subjects. In both cases, it is possible to compute accident rates for certain driver-vehicle-road-environment classes. Because of discrepancies between the two sources of accident data, the accident rates would not be the same in many of the classes. Future programs to improve or correct various sources of accident data

may resolve the discrepancies, but the high cost of the programs make them unfeasible in the near future. Therefore it is useful to compare the alternative sources of accident-involvement data which might be used in the near future for calculation of accident rates of the various classes.

If an annual exposure survey is conducted with a sample of drivers in a given jurisdiction, three alternative sources of accident-involvement data may be considered for calculation of accident rates: the mass accident data for the jurisdiction in the given year; estimates of accident involvement by the subjects of the exposure survey for the given year; and the official state accident records of the individual subjects in the survey for the given year.

Mass accident data is available from most states on an annual basis, and there is a continuing trend for states to put the data on computer tapes for convenient analysis. Although the data is biased due to underreporting of accidents, many states are striving to reduce the biases by improved reporting procedures. At present, the mass accident data of official state records are the best available source. All reported accident involvements may be counted from the tapes to determine frequencies of accident involvements in each driver-vehicle-road-environment class. Even though each driver-vehicle involvement in a multi-vehicle accident may be in a different class, the frequency of accidents may be classified by identifying each accident according to the class of the at-fault driver-vehicle combination. The costs of classifying accidents and involvements would be minimal, assuming that the necessary variables are included in the state data tapes. Both accident rates and accident-involvement rates would be determined by dividing frequencies in classes by corresponding exposure values.

Estimates of accident involvements in a calendar year by subjects in exposure surveys would provide rather small samples of accident involvements. It would be desirable for each involve-

ment to be classified according to vehicle, road and environment by the subject, even though it might be difficult to recall after a year's time. (In the pilot survey of Volume I, the subjects did not indicate the classifications of their involvements.) It was determined in Volume II that the accident involvements admitted in the pilot-survey were only about one-third of all accidents actually experienced by the subjects. Procedures would be needed in future surveys to reduce the biases due to this underreporting. However, the requirement for classifying admitted accidents might have an effect of even worse underreporting. Advantages of this method are the minimal effort required to obtain accident data, and the fact that all data (exposure and accident data) could be placed on one tape. Involvement frequencies and exposure totals could be computed separately, by class, and combined in group accident rates. Also, individual accident-involvement rates for each driver (or each vehicle-road-environment class of each driver) could be computed if it was desired to obtain mean accident involvement rates in an AID analysis.

Finally, official state accident records of exposure survey subjects could be obtained by record searches after the survey, and a separate accident-involvement file could be created. If desired, it could be merged with the exposure file, by subject. As in the previous alternative, underreporting biases would be present in the accident-involvement data. The extra costs of the record search would be substantial, with no known advantage over the other two methods. However, a combination of this method with admitted involvements of subjects in the previous method would provide a partial means of eliminating underreporting biases.

It is clear that the first method (use of mass accident data) is superior to the others because of its nominal cost, larger sample size, and greater potential for eliminating biases in accident rates due to underreporting of accidents. However, it would

be useful to ask for accident-involvement estimates in future exposure surveys (as in the second method above) for use in continued analysis of underreporting biases.

COMBINED DATA FROM INDEPENDENT EXPOSURE SOURCES

In Volume I, it was concluded that licensed drivers, sampled randomly, would be the best source of exposure data. Further, it was concluded that one-day trip logs of miles driven in various vehicle-road-environment combinations would be the best way for survey subjects to provide estimates of their individual exposure. It was assumed that the accurate estimates of exposure in the final classifications of exposure (i.e. driver-vehicle-road-environment classes) could be best obtained in a single, comprehensive survey in which all driving was classified by all of the independent variables needed for classification. It was later realized that complete classifications of exposure might be obtained by combining several independent sources of exposure information. One reason for this consideration is the fact that some exposure data is currently available on a continuing basis, even though it is not fully classified (e.g. annual statewide mileage estimates, classified by road type only, are estimated on the basis of gasoline consumption in each state, coupled with limited traffic counts). Another reason for considering the combination of independent exposure sources instead of a single survey is the fact that each independent exposure source could consider a fairly small number of related variables, and thus could perhaps achieve greater accuracy. Survey subjects might be more likely to respond if the survey was quite short, and their responses might be more accurate. Other exposure sources (e.g. roadside observations) might be more accurate than assumed in Volume I if the number of variables was quite small.

An example of the potential for combining independent exposure sources is available from the data file used in Section 2.

The independent variables were divided into three groups, as shown in Table 4, and separate AID hierarchies of exposure classes were produced for the three variable groups, as shown in Figures 6, 7 and 8. Each of these hierarchies can be considered as if it were produced from an exposure data source which was independent of the other two. For example, each could be assumed to be the result of independent surveys, or they could be assumed to come from three different kinds of sources, such as one driver survey, one roadside observation study and one gasoline consumption analysis.

If a combined hierarchy of driver-vehicle-road-environment classes is supposed to include the two most important variables from each of the three independent sources, it would include cross-classifications among the three sets of variables, similar to the hierarchies of Sections 2 and 3. But it is clear that the hierarchies of Figures 6, 7 and 8 do not provide the necessary information (interaction among the three sets of variables) to produce such cross-classifications. Therefore, independent sources of exposure data must include some common independent variables in order to define the necessary cross-classifications. At the very least, each independent source must include the variable which is the first splitting variable in a combined, total hierarchy.

Probably the most useful example of this principle can be seen with respect tothe recommended hierarchy of Section 3 (see Figure 5.) If exposure values are to be determined for each of its classes by means of two or more independent data sources, then each source must include Driver Sex as one of its independent variables. Further, from the second splitting level on the hierarchy of Figure 5 it can be seen that the exposure of every male in the data source must be classified as to whether it was in a passenger car or in another vehicle type. At this point, it is possible to identify three potential sources for independent exposure data: females, males in passenger cars, and males in other vehicles. Whatever the sources, these three groups must be identified. Clearly, any source that does not identify Driver Sex

Table 4

Groupings of Independent Variables
for Separate Classification Analysis

Driver Variables

- Driver Sex
- Driver Age

Vehicle Variables

- Vehicle Type
- Model Year
- Passenger Car Size
- Passenger Car Make

Road-Environment Variables

- Percent Driving on Streets
- Percent Driving on Urban Freeways
- Percent Driving on Rural Freeways
- Percent Driving on Rural Roads
- Percent Driving at Night
- Percent Driving on Wet Roadway

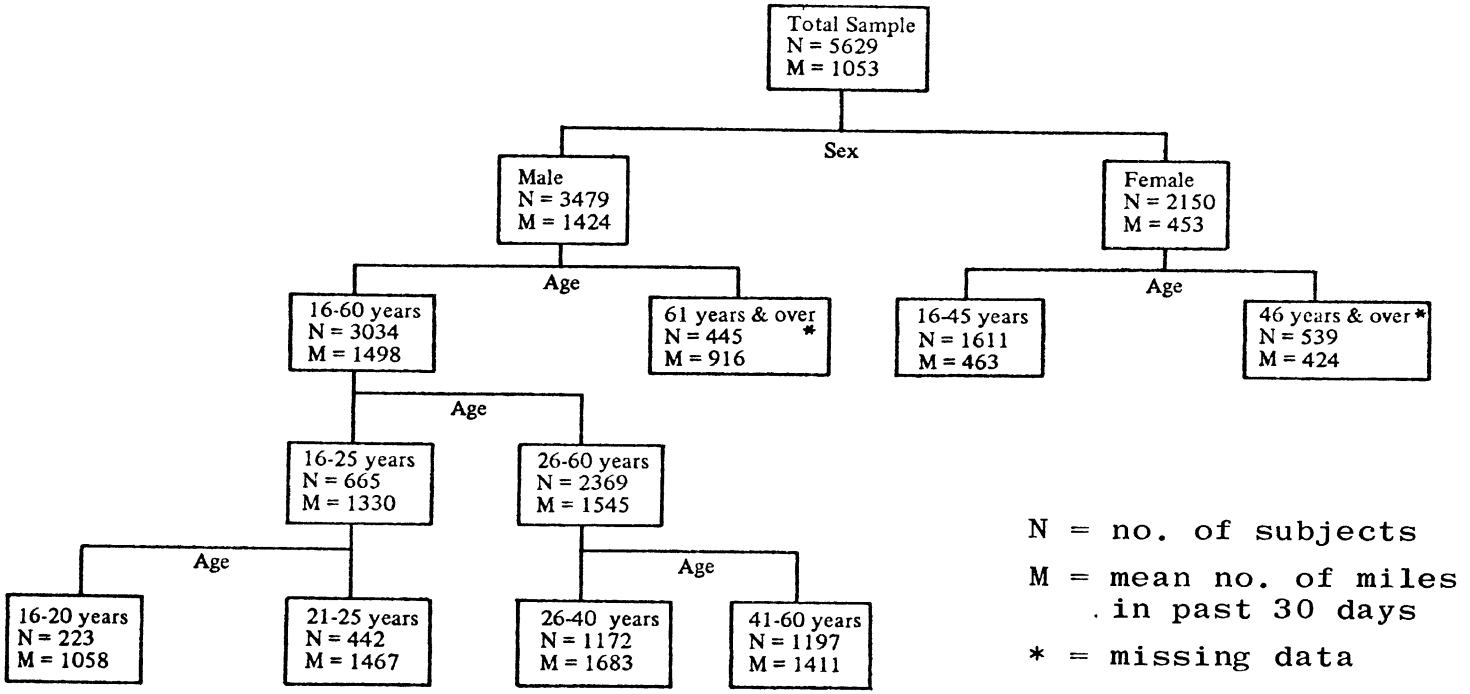


Figure 6 -- Driver Exposure Classes

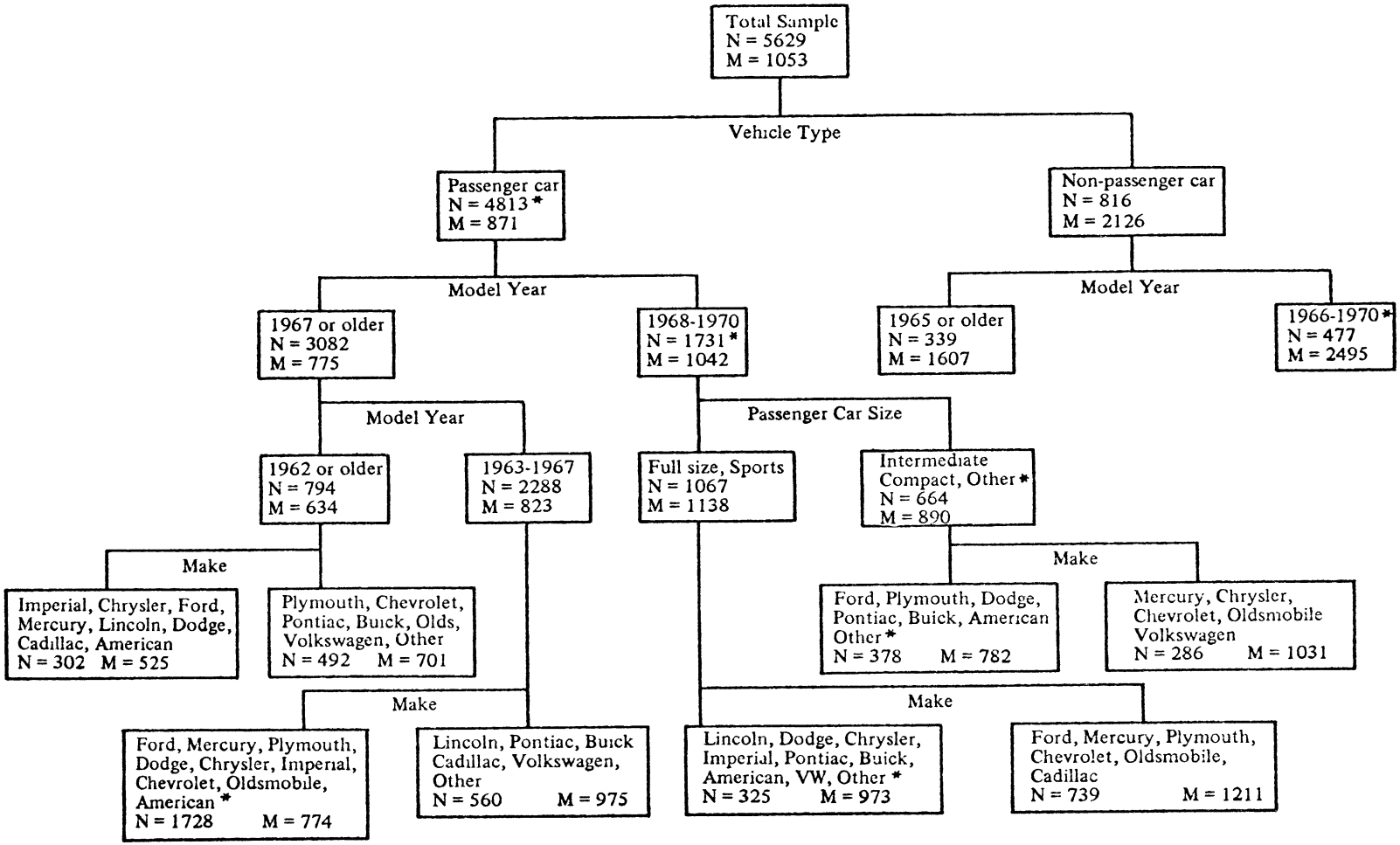


Figure 7 -- Vehicle Exposure Classes

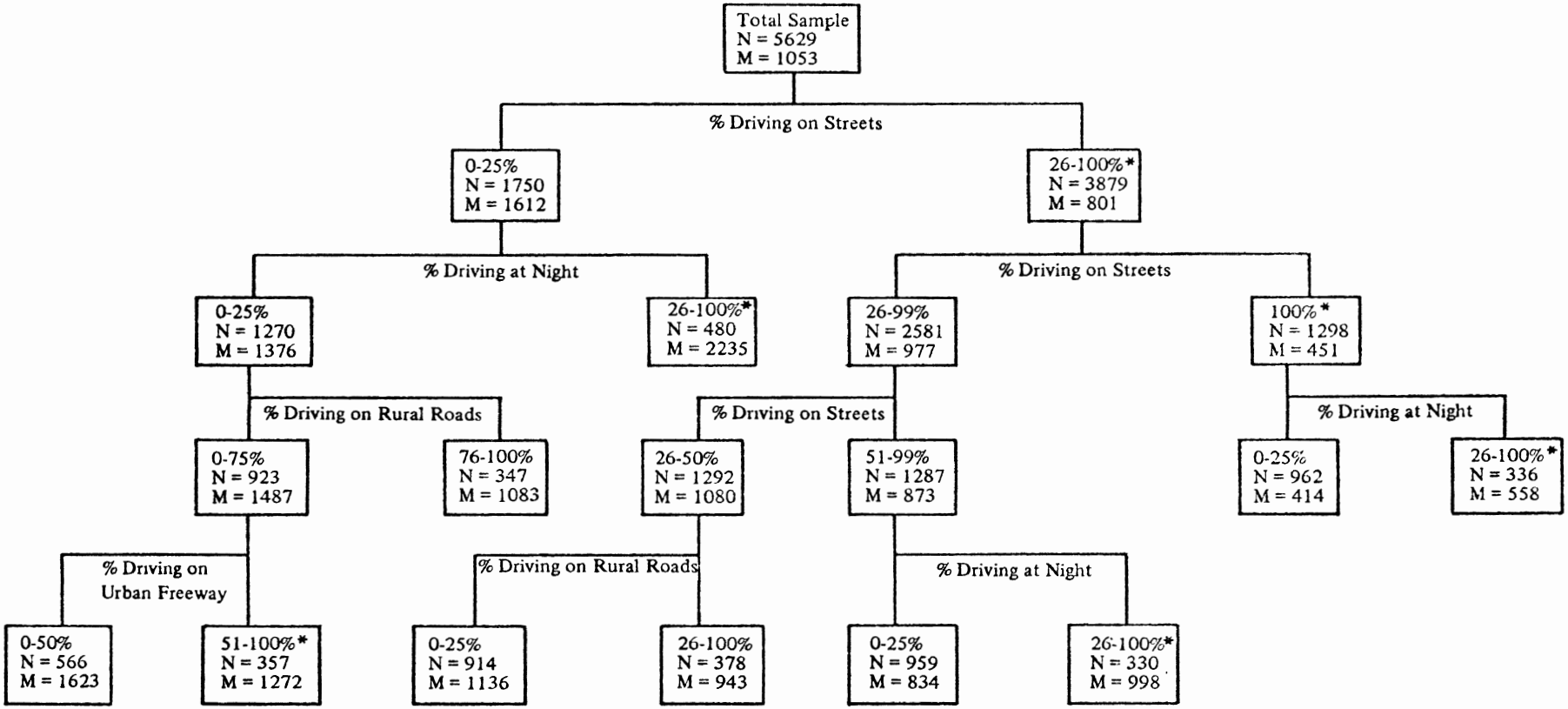


Figure 8 -- Road-Environment Exposure Classes

may not be used; unfortunately, this restriction eliminates the current sources of exposure data, by road type, from analysis of gasoline consumption. Similarly, Driver Age must be obtained for a great majority of the exposure data.

Two of the variables in Figure 5 - Sex and Road Type - are required for all of the classes. Also, the variables Age and Vehicle Type are required for about 80% of the total exposure. Thus, it appears that a single, comprehensive exposure survey of drivers is the best approach for obtaining data in all of the classes above the final splitting level. However, because the Day/Night and Model Year variables apply to only a few classes at the bottom level, another data-collection approach may be desirable at that point.

In order to obtain Model Year classes, a special survey of owners of non-passenger-car vehicles could be made. The sample could be obtained from state vehicle registration records. Subjects would be asked the model year of their vehicles and estimates of mileage driven by males on streets and on other roads. A problem with this method could arise if the exposure values of the two methods did not check for the two groups above the Model Year splitting level (i.e. Males, Other Vehicles, Streets and Males, Other Vehicles, Other Roads). However, the added expense of the special survey might be worth the potential improved accuracy of the main survey due to not asking Model Year. But perhaps the same result could be achieved in the main survey by designing the questionnaire so that only those drivers with exposure in Other Vehicles (non-passenger-cars) would be bothered with a request for Model Year information.

In order to obtain Day/Night classes, a special survey of 26-40 year old males could be made. The sample could be obtained from proper strata of state driver lists. Subjects would be asked for estimates of passenger car driving in four classes: streets/day,

streets/night, other roads/day, and other roads/night. As in the paragraph above, this method could involve a problem if the exposure values of the groups above the Day/Night splits did not check with values from the main survey. In this case, the special survey may not be worth its extra cost because the day/night division of exposure would not be an especially difficult task for most of the subjects in the main survey.

The result of this analysis is a confirmation of the previously recommended concept of a single, comprehensive exposure survey, with the possible exception of a special survey for Model Year data of non-passenger-cars. This result does not offer the potential of cost savings by the use of existing exposure-data sources. Any attempt to link two or more independent sources of direct exposure data would require commonality of most of the variables in each source, and hence would add to the cost if required sample size were adhered to.

The use of induced exposure data may also be considered as a means of estimating the exposure in some of the classes at the lower levels of the hierarchy. If direct exposure measurements are known for the second lowest level, accident data may be used to determine the distributions of direct exposure between the two lowest classes which split a class at the second lowest level. The induced exposure method proposed by Thorpe* may be used for this purpose. In the hierarchy of Figure 5, this method would involve the identification of 18 classes of accident involvements, but it would not require the knowledge of which driver-vehicle combination was at fault in a multi-vehicle accident. In the future, when at-fault drivers are accurately identified in mass accident data, the method proposed by Haight[1] could be used. By either method, rough estimates of the four Day/Night classes and the four Model Year classes of the Figure 5 hierarchy could be made.

---

* Reference 7 of Volume I.

[1] Haight, F.A., A Crude Framework for Bypassing Exposure, Journal of Safety Research, March, 1970.

One problem with these procedures is the possibility that the distributions of exposure among the 14 classes at the second lowest level would not check when the direct exposure and induced exposure methods are compared. Further research should be performed on such comparisons.

SECTION 5
CONCLUSIONS AND RECOMMENDATIONS

1. A group accident-rate difference method should be used to establish homogeneous driver-vehicle-road-environment classes for the determination of exposure and accident involvement, and for the calculation of accident rates which apply to evaluation of highway safety countermeasures. The group accident-rate difference method splits a set of data cases according to the successive independent variables which produce maximum relative differences in group accident rate, thus creating a hierarchy of unique accident-rate classes.

2. The following independent variables are recommended for use in future exposure surveys: Driver Sex, Driver Age, Vehicle Type, Model Year, Vehicle Make, Road Type, Day/Night.

3. The hierarchy of Figure 5 is recommended for the calculation of accident rates resulting from the combination of mass accident data and exposure survey data. The hierarchy contains 18 unique classes of driver-vehicle-road-environment combinations, produced by the group accident-rate difference method, with slight modifications for purposes of symmetry and uniformity in class size.

4. Mass accident data from official state records should be used to determine accident-involvement frequencies in the classes of the accident-rate hierarchy. Accident data collected for the subjects of an exposure survey would be disadvantageous to the accurate determination of class accident rates.

5. Exposure data for use in determination of class accident rates should be collected in a single, comprehensive survey, rather than by combination of independent sources of exposure data.

6. Future research should be performed to determine the accuracy of induced exposure calculations for some of the lower levels in the accident-rate hierarchy.

APPENDIX N

SUPPLEMENTARY STATEMENT OF WORK

1. Using existing data from a national exposure survey, determine a hierarchy of exposure classifications based on individual accident rates as the dependent variable, using an AID-analysis methodology.

2. Determine the average accident rates, by class, in the hierarchy found in Task 1.

3. Establish an alternative methodology for determining homogeneous exposure classifications based on accident rates, by means of a step-by-step analogy to AID analysis which uses average accident rates by class instead of individual accident rates.

4. Using the same data as in Task 1, and the methodology of Task 3, determine an alternative hierarchy of homogeneous exposure classifications based on accident rates.

5. Compare the two hierarchies of exposure classification in terms of structure and accident-rate values and interpret their differences with respect to statistical significance.

6. Recommend and support the preferred analytical methodology for selecting homogeneous exposure classifications based on accident data.

7. Compare the structures of the preferred accident-rate-based classifications and the exposure-based classifications of the previous study. If appropriate, recommend changes in previous recommendations for independent variables to be included in exposure surveys.

8. Describe methodologies for deriving accident rates within exposure classifications in three situations: where accident data is determined from individuals in an exposure survey; where accident data is determined from official records of individuals in an exposure survey; and where mass accident data is used from the jurisdictions covered in an exposure survey.

9. Using the same data as in Task 1, perform three AID analyses on driver variables, vehicle variables, and roadway variables separately. Investigate procedures for combining the three sets of data classifications as if they had come from independent sources.

APPENDIX O
EXPOSURE AND ACCIDENT DISTRIBUTIONS

Miles vs. No. of Accidents

Accidents in 3 Years	Number of Subjects	Mean Miles in 30 days
0	4327	989
1	1102	1168
2	210	1110
3	54	1427
4	19	2034
5	6	1583
6	1	250
7	0	--
8	1	3440
Total	5720	1036

Miles and Accidents vs. Vehicle Type

Type	Number	Mean Miles	Mean No. Acc.
Passenger Car	4825	866	.325
Small Truck	626	1536	.225
Large Truck	60	4035	.317
Tr-Trailer	59	5543	.390
Taxi	9	5444	.556
Bus	25	3057	.600
Other	39	2026	.256
Total	5643	1048	.315

Miles and Accidents vs. Passenger Car Size

Size	Number	Mean Miles	Mean No. Acc.
Full	3020	886	.296
Inter	857	772	.340
Compact	726	821	.410
Sports	152	1160	.408
Other	39	1066	.333
Total	4794	865	.324

Miles and Accidents vs. Sex

Sex	Number	Mean Miles	Mean No. Acc.
Male	3520	1407	.376
Female	2200	443	.219
Total	5720	1036	.315

Miles and Accidents vs. % Drive Night

%	Number	Mean Miles	Mean No. Acc.
0	1515	648	.220
1-25	2599	1117	.294
26-50	1110	1435	.416
51-76	182	1375	.654
76-99	115	1190	.522
100	24	704	.583
Total	5545	1060	.316

Miles and Accidents vs. % Drive Wet

%	Number	Mean Miles	Mean No. Acc.
0	2994	864	.279
1-25	1480	1251	.364
26-50	812	1404	.340
51-75	103	1211	.466
76-99	50	1010	.420
100	41	1360	.268
Total	5480	1060	.316

Miles and Accidents vs. % Drive Streets

%	Number	Mean Miles	Mean No. Acc.
0	399	1455	.296
1-25	1374	1631	.354
26-50	1294	1078	.347
51-75	561	939	.351
76-99	729	822	.347
100	1159	443	.211
Total	5516	1061	.317

Miles and Accidents vs. % Drive Urban Freeway

%	Number	Mean Miles	Mean No. Acc.
0	2735	817	.268
1-25	1297	1278	.345
26-50	864	1196	.412
51-75	290	1569	.383
76-99	310	1438	.300
100	19	1483	.368
Total	5515	1061	.317

Miles and Accidents vs. % Drive Rural Freeway

%	Number	Mean Miles	Mean No. Acc.
0	4049	855	.289
1-25	711	1535	.398
26-50	474	1549	.397
51-75	157	1885	.363
76-99	110	2171	.418
100	13	2425	.308
Total	5514	1061	.317

Miles and Accidents vs. % Drive Rural Roads

%	Number	Mean Miles	Mean No. Acc.
0	3555	893	.307
1-25	818	1518	.414
26-50	538	1331	.320
51-75	177	1371	.316
76-99	264	1172	.227
100	163	1031	.190
Total	5515	1061	.317

Miles and Accidents vs. Model Year

Year	Number	Mean Miles	Mean No. Acc.
1960 or older	523	729	.241
1961	161	852	.230
1962	285	780	.256
1963	407	914	.295
1964	434	925	.288
1965	571	910	.320
1966	614	1063	.314
1967	623	1013	.326
1968	738	1168	.314
1969	803	1225	.366
1970	461	1563	.401
Total	5620	1047	.315

Miles and Accidents vs. Age

Years	Number	Mean Miles	Mean No. Acc.
16-20	372	812	.707
21-25	754	1043	.511
26-30	701	1190	.371
31-35	619	1163	.323
36-40	582	1282	.234
41-45	599	1035	.214
46-50	569	1059	.250
51-60	844	984	.180
61-70	492	721	.171
71 and over	136	417	.221
Total	5668	1033	.314

Miles and Accidents vs. Make

Make	Number	Mean Miles	Mean No. Acc.
Ford	1048	824	.323
Mercury	172	883	.262
Lincoln	22	808	.273
Plymouth	298	847	.332
Dodge	254	808	.366
Chrysler	114	797	.211
Imperial	5	294	0
Chevrolet	1191	895	.320
Pontiac	346	920	.338
Buick	346	920	.338
Oldsmobile	289	934	.336
Cadillac	116	926	.181
American	175	643	.211
Volkswagen	256	955	.441
Other	192	903	.443
Total	4793	866	.324