Phase II

EVALUATION OF THE NATIONAL CRASH EXPERIENCE:

COMPARISON OF CARDfile NATIONAL MOTOR VEHICLE ACCIDENT PROJECTIONS WITH PROJECTIONS FROM OTHER DATA BASES

BY

SANTO SALVATORE

ROBERT WALTER

PETER MENGERT

A. INTRODUCTION

1. Purpose

The purpose of this study is to compare national motor vehicle accident projections made from the Crash Avoidance Research Data base (CARDfile) with national motor vehicle accident projections made from other data bases. For the most part, the comparison will be with data derived from the computerized data bank of the National Accident Sampling System (NASS). Where appropriate, data from the National Safety Council's "Accident Facts" will also be utilized. Phase I of this project compared the distribution of people, vehicles and roads in the CARDfile states with the nation.

2. Description of CARDfile and NASS

CARDfile consists of crash data extracted from the automated police accident reports of the following six states:

* Indiana
* Maryland
* Michigan
* Pennsylvania
* Texas
* Washington

CARDfile has been designed to: (1) provide the information needed to examine the relationship between selected crash avoidance vehicle design characteristics and their crash propensity; and (2) to support problem identification activities in crash avoidance research. CARDfile now contains in excess of 4.0 million accidents involving 6.5 million vehicles covering the crash experience of the latest three years, 1983-1985, (Edwards, 1987). States were selected for inclusion in CARDfile on the basis of data availability in machine-readable form and commonality of data fields as well as population characteristics.

NASS investigates in depth a sample of police accident reports in a three stage, unequal probability selection plan. The nation is first broken down into 1279 probability sampling units(PSUs) which are categorized into 70 strata. Some PSUs are selected with certainty and others with probability proportional to their 1977 population. The second stage samples police jurisdictions within the PSUs with the large jurisdictions being oversampled. Stage 3 selects the accidents in a fashion that oversamples the more severe and the more rare type of accident.

For the comparison year, 1984, NASS is based on 11,598 accidents. National accident estimates are computed by multiplying the number of accidents sampled in a subcategory by the ratio inflation factor. The ratio inflation factor is the inverse of the corrected probability of selecting the accident type.

B. METHODOLOGY

1. Importance of Variable Matching

An important aspect of this analysis is the detailed mapping of the variables in NASS to the variables in CARDfile. This is most important since improper matching of

variables values in the two data bases will result in discrepancies which are not due to the crash experience. A general, preliminary mapping of variables in CARDfile with NASS was developed by TSC and refined by NHTSA's Crash Avoidance contractor, Mr. James MacDonough. The mapping of the NASS values onto CARDfile values for the accident file variable WEATHER is shown below. WEATHER is one of the simpler variables to be considered. Mapping of all the variables in the analysis is provided in Appendix 1.

CARDfile VALUES	CORRESPONDING NASS VALUES
1 - MISSING	
2 - UNKNOWN	9 - UNKNOWN
3 - CLEAR/CLOUDY	1 - NO ADVERSE WEATHER
4 - RAIN	2,6 - RAIN, RAIN + FOG
5 - SNOW/ICE	3,4,7 - SLEET, SNOW, SLEET + FOG
6 - OTHER	5,8 - FOG, SMOG, SMOKE, ETC.

2. Selection of Variables for Comparison

The variables for comparison were chosen by NHTSA on the bases of dual criteria. The variables chosen were: (1) descriptive of the crash; and (2) comparable to data available from Phase I results. The selected variables, arranged in accordance to their appearance in the CARDfile data base file, are shown below.

ACCIDENT FILE
- Light Conditions
- Weather Conditions
- Primary Impact
- Relation to Intersection
- Intersection Characteristics
- Number of Vehicles Involved

VEHICLE FILE
- Vehicle Type
- Model Year by Vehicle Type
- Component Failure by Vehicle Type

DRIVER FILE
- Sex by Age

3. Data Collection and Presentation

Computer programs were then developed by TSC to aggregate the NASS values in correspondance to the CARDfile classification scheme. The univariate distributions in the corresponding categories for CARDfile were obtained from NHTSA; the bivariate distributions in the vehicle and driver files were extracted from CARDfile by TSC. The data were then collated into tabular format and indices comparing the two data bases derived using personal computer software.

The method of presentation will be similar to that used in the report on Phase I of this project and as shown in the sample table below. The body of this report refers to tabular data given in tables 1 through 10 collected in Appendix 2. These tables

provide summary statistics comparing the CARDfile and NASS results. Corresponding data for the six individual states are included in Appendix 3. Appendix 3 may be examined to discover which states may be driving the predictions.

WEATHER CONDITIONS (SAMPLE TABLE)

(1) Weather Condition	(2) CARDfile States Actual	(3) CARD Predicted Nation	(4) Percent of Accident Involved Drivers	(5) NASS Predicted Nation	(6) Percent of Accident Involved Drivers	(7) Ratio CARD% NASS%
Clear/Cloudy	1506840	4808622	80.02	4638911	79.63	1.00
Rain	186663	849316	14.13	789350	13.55	1.04
Snow/Ice	53024	241259	4.01	340627	5.85	0.69
Other	17942	81636	1.36	13878	0.24	5.70
Missing/Unk	6240	28392	0.47	42802	0.73	0.64
Total	1320709	6009225	100.00	5825568	100.00	1.03

Each table consists of seven columns. Column 1 lists subclasses or values of the CARDfile characteristic or variable under consideration such as age brackets, type of weather, number of vehicles in the accident.

Column 2 provides the actual number of accidents found in CARDfile for that particular value or subclass. Column 3 is the number of accidents in the nation predicted from the CARDfile sample of accidents. This prediction is based on the simple multiplier of 4.55, i.e., the inverse of .22, the proportion representing, approximately, the national population, the licensed driver population, and the vehicle population in the CARDfile states. This assumes that the national crash experience is a simple multiple of these populations. (The use of a multiplier in this vicinity is substantiated by the fatality count obtained in Phase I. There it was found, from the Fatal Accident Reporting System (FARS), that the CARDfile states suffered 9483 of the nations 44250 fatalities. This is 21.43% of the fatalities which would provide a multiplier of 4.67.) Column 4 is the percentage of the predicted accidents falling in the subclass.

Column 5 and 6 provide similar information for the NASS derived statistics. Column 5 is the number of accidents in the nation predicted by the NASS sample and Column 6 the percentage of the predicted accidents falling in the subclass.

Column 7 is an index number which may be used for comparative purposes. For the Total this number is the ratio of the number of accidents or drivers predicted by CARDfile to the number of accidents predicted by NASS. For the individual values, such as rain, this index number is the ratio of the two data bases percentages as shown in Column 4 and Column 6.

C. RESULTS

General Comment

A glance at Table 1 reveals that both CARDfile and NASS predict approximately six million accidents occurred in the nation in 1984. They differ from each other by slightly over 100,000 accidents. The ratio of CARDfile to NASS is 1.02. This index may be interpreted to say that CARDfile predicts two percent more accidents than NASS or that NASS predicts two percent less accidents than CARDfile.

It should be noted that for all the tables the total number of predicted accidents is the same; the composition of the subcategories will differ and provide predictive differences between the two data bases.

Although in most of the important subcategories of the variables there is excellent agreement between CARDfile and NASS, predicted differences in the vicinity of + or — 10% or more will be pointed out.Smaller differences will be pointed out for categories with a large number of accidents and bigger differences for categories with a small number of accidents. An attempt will be made to relate differences to Phase I results of this project. The purpose here is to document the extent to which over- and under-prediction by CARDfile in relation to NASS are correlated to over- and under-representation of CARDfile population characteristics as compared to the nation in Phase I.

Such a correlation would provide validity of a type. Cursory comparisons to other sources of data will also be made to further this validation process.

The representativeness or validity of the national crash experience predicted by CARDfile will be taken up in the discussion section of this paper.

TABLE 1 LIGHT CONDITIONS

For the majority of the accidents, which occur during the day (62%) there is excellent agreement between the two data bases. CARDfile predicts more accidents to occur at night on unlighted roads and at dawn than NASS while predicting fewer accidents to occur on dark, lighted roads and at dusk. This writer sees no obvious explanation for the result that ties it to Phase I results. It should be pointed out that the definitions of dusk, dawn, and night lighted are not clear and somewhat subjective. It is of interest that the sums of dark lighted and dark are approximately equal in CARDfile (33.04%) and NASS (32.53%).

TABLE 2 WEATHER CONDITIONS

Again, for the large majority of accidents, 80%, which occur in clear/cloudy conditions, there is excellent agreement between the two data bases. CARDfile predicts fewer accidents on roads with ice and snow; considerably more accidents in the "other" category which includes fog, smog, and smoke, and slightly more accidents in rain. However, the "other" category is small for both CARDfile and NASS (1.36% and 0.24% respectively).

A cursory examination of climatic conditions in the CARDfile states as compared to the nation, indicates that these over- and under-predictions of accidents nationwide

may reflect the CARDfile states climate. Again, for the large majority of accidents, 80%, which occur in clear/cloudy conditions there is excellent agreement between the data bases.

TABLE 3 PRIMARY IMPACT

In the principal category of vehicle in transport, (64%) CARDfile and NASS are in excellent agreement with a ratio of 1.00.

CARDfile predicts more accidents with parked cars(vehicles not in transport), trains and of the rollover type than NASS. That CARDfile predicts more accidents with parked cars is congruent with the observation made in Phase I that the CARDfile states have more urban roadway and therefore more parked vehicles.

Other sources (the National Atlas of the United States of America, U.S. Department of the Interior Geological Survey, 1970, Washington, DC) indicate that there is a greater proportion of railroad mileage in the CARDfile states than the rest of the nation and, therefore, it is expected that CARDfile would overpredict motor vehicle accidents with trains. However, in both data bases train accidents are a very small portion (~0.1%) of the total accidents.

CARDfile predicts fewer accidents with pedalists and accidents of the non-crash variety than NASS.

TABLE 4 RELATION TO INTERSECTION

Accidents not intersection related are projected with good agreement (CARDfile 47.11%, NASS 47.94%).

CARDfile predicts more accidents as being intersection related than does NASS and predicts fewer accidents related to driveways. This result is in agreement with Phase I results. Those results indicate that the CARDfile states have more urban roadways than the nation thus possibly having more intersections and fewer driveways.

TABLE 5 INTERSECTION CHARACTERISTICS

Table 5 analyzes the intersection accidents. Of those accidents occurring at the intersection, CARDfile predicts considerably more to take place at intersections with stop signs or yield signs. This would again be congruent with an urban environment: more accidents at intersections, more accidents at intersections with signs.

TABLE 6 NUMBER OF VEHICLES

CARDfile predicts fewer single vehicle accidents than does NASS. CARDfile also predicts considerably more accidents with 3 or more vehicles than does NASS (6.45% vs 4.58%). These results agree with Table 3 results which show fewer noncrash type of accidents and accidents with parked cars. The greater number of multiple car accidents may also be expected in an environment with more urban roads.

TABLE 7 VEHICLE TYPE

CARDfile predicts slightly less accidents with passenger cars which is consistent with Phase I results that show the CARDfile states to have fewer registered cars. However, the Phase I results showed that light trucks were overrepresented in the CARDfile states but slightly underrepresented in the accident data shown here.

CARDfile predicts a greater percentage of medium and heavy trucks to be involved in accidents. This is congruent with Phase I results which showed a greater percentage of registered medium and heavy trucks in the CARDfile states.

TABLE 8 VEHICLE TYPE BY MODEL YEAR

CARDfile predicts less passenger car accidents in 1982 and more passenger car accidents for 1981 models than does NASS. CARDfile also predicts more light truck accidents for 1982, 1981 and 1975-1979 models and fewer light truck accidents for 1965-1969 models. No comparable data were developed during Phase I.

TABLE 9 VEHICLE TYPE BY DEFECT

Over 98 percent of the vehicles are found not to have a defect or are unknown to have a defect in both CARDfile and NASS. CARDfile predicts more defects of all types for passenger cars and light trucks than NASS(1.90% vs 1.34%). The numbers with which we are dealing here are, however, quite small. Still the ratios for the most part are not tremendously variable indicating a core of stability.

In both data bases, brakes, tires and steering are ranked as the most common defects.

TABLE 10A DRIVERS IN ACCIDENTS BY SEX AND AGE

This table has redistributed drivers with unknown age/sex according to the existing distribution of age and sex. (Table 10B shows the original driver distribution as originally output from the computer file.) In the overall age category CARDfile overpredicts the 30-34 and 50-54 age brackets and underpredicts the over 70 age bracket.

Broken down by sex CARDfile overpredicts females in accidents in the 50-54 and 65-69 age brackets and underpredicts females in accidents in the 55-59, 60-64 and over 70 age bracket.

CARDfile overpredicts males in accidents in the 25-29, 30-34, 40-44, 50-54 age brackets and underpredicts males in the 65-69 and over 70 age brackets. The relationship of these results to Phase I results is not clear. This fluctuation suggests "noise." The noise is probably due to NASS since the CARDfile sample are relatively large.

D. DISCUSSION

1. The Problem

We have compared national projections of motor vehicle accidents based on CARDfile and NASS. The total number of accidents predicted from the two data bases, the number of accidents predicted in large categories and the number of accidents predicted in many small categories are in very good agreement. The question to be answered is "What is the significance of this agreement?" Also to be answered is the question "What is the significance of disagreement when it occurs?"

Phase I of this project compared a number of population characteristics of the CARDfile states with the nation. Variables included age and sex distribution of licensed drivers; distribution of vehicles by type, model years, make and size; distribution of road types and vehicle miles traveled. Results showed that the population characteristics of the CARDfile states, with minor differences, mirror the nation closely. It was concluded that though the CARDfile states were not chosen at random their population characteristics typify the nation.
~~The current~~ Phase II is an effort to determine the degree to which crash data elements in CARDfile typify the national crash experience.

Generally, the national crash experience is known within uncertain boundaries. (An exception is fatal accidents for which a census is maintained in FARS).

NASS is the most comprehensive attempt available to construct the national crash experience on the basis of a randomly selected sample of accidents. For that reason NASS was chosen for comparison. In essence, we compare two estimates of the national crash experience for each crash data element.

2. Estimates of the Total Number of Accidents

The national crash experience may be summarized by a number representing all the accidents in the nation. CARDfile and NASS both predict close to 6 million vehicles to have occurred nationwide in 1984. It might be concluded that this agreement signifies mutual corroboration of the two data bases. However an examination of "Accident Facts" shows that the National Safety Council estimated 18 million accidents to have occurred nationally in 1984. Obviously different criteria ~~of~~ for accidents have been used.

The criteria can change the accident data bases in two ways. One, the number of total accidents will vary with the criteria. Two, the criteria will change the composition of the values within the variable.

In both CARDfile and NASS severe accidents are more likely to be included. Within CARDfile all accidents as severe as tow-aways are included. However, some of the CARDfile states have a reporting threshold as low as $200.00 property damage. In NASS, the more severe and rare accidents are oversampled relative to less severe and more common accidents.

Fair representation of the nation's crash experience in severe accidents is therefore likely. However, the more numerous, minor accidents will, as a rule, not be as likely to be included in the data base resulting in a possible distortion of the values

comprising the variables. In comparing the two data bases projections this point should be kept in mind as a possible explanatory principle.

3. Agreements and Differences Between the Two Data Bases

The two data bases predict approximately equal proportions of accidents in many categories. These agreements include accidents occuring during the day (62.49% CARDfile, 62.60% NASS); accidents occurring in clear/cloudy weather conditions (80.08% CARDfile, 79.63% NASS); accidents involving vehicles in transport (64.57% CARDfile, 64.72% NASS); accidents involved vehicles without defects (98.10% CARDfile, 98.61% NASS).

CARDfile predicts a greater proportion of accidents than NASS: (1) to have primary impact with parked cars; (2) to occur at intersections; (3) are likely to have signs; and (4) to involve 3 or more vehicles. These accident characteristics appear to describe urban type accidents. This result is congruent with Phase I results which found the CARDfile states to have more urban road mileage than the rest of the nation. This appears to be comfirmation of CARDfile's ability to stably reflect accident parameters tied down to population characteristics. These correlations may point up possible correction factors to the projections.
Two additional elements of disagreement may be considered in developing further this rational argument for CARDfile validity. CARDfile overpredicts accidents with trains (.13% vs .07%) and under predicts accidents under conditions of ice and snow (4.01% vs 5.85%). These latter two factors appear to be definitely tied into more railroad track mileage and smaller snowfall in the CARDfile states.

4. The Sampling Problem

Both CARDfile and NASS project the national crash experience on the basis of a sample of accidents. It is probably safe to say that all national motor vehicle crash estimates (outside of the fatal accidents in FARS) are based on a sample and not a census.

The CARDfile sample is 1,321,000 accidents for 1984. The corresponding number for NASS is 11,598. Questions of validity or fidelity of the estimates aside, the estimate based on the larger sample will be more statistically stable. See Addendum for a discussion of the computation of the standard error for CARDfile and NASS.

However, the CARDfile states and the accidents therein were not chosen randomly but on the basis of data availability, geographic representation of the nation, and admnistrative factors. It can then be argued, on the other hand, that the predictions based on CARDfile will be biased. Phase I of this project demonstrated that the CARDfile states' population characteristics are mostly typical of the nation. Deviation of population characteristics in Phase I have been found to correlate to over- and under-prediction of accident characteristics in this current analysis. The indication is that estimates of the national crash experience derived from CARDfile may need to be corrected for CARDfile states population characteristics.

The general statistical rule states that the larger the sample size the more reliable or statistically stable is the prediction of the true population value - in this case the national crash experience. The obverse of this is that the smaller sample size will result in greater random fluctuations in the prediction of the true population value.

Obviously any projection made from CARDfile, for any comparable characteristic, will be based on a larger sample than the projection from NASS. Additionally, for both data bases, the sample size varies with the value selected in the variable. For example the projection of the number of accidents occurring during the day will be based on a much larger sample than projections of accidents occurring at dawn. The implication is that this reduction in sample size will create a greater sampling fluctuation in NASS projections than CARDfile projections because the NASS sample is relatively small to begin with.

We previously mentioned the apparent "noise" in the data relating to age and sex of the driver

Figure 11A plots the ratio % of accident involved drivers to licensed drivers (data from Phase I) for each age group. The % accident involved drivers are predictions derived from CARDfile and NASS. The line function based on CARDfile appears to be smoother.

Figures 11A and 11B provide the same ratio for the female and male populations respectively. However, the numerator here is the percent of the population in the age group since a breakdown of licensed drivers by age and sex was not located.

From the above we deduce that agreement between the two data bases is more likely in subcategories which have a considerable proportion of accidents. Obversely, disagreement is more likely in subcategories which have a small proportion of accidents. Further, the disagreement is more likely to be due to NASS sampling fluctuations because of its smaller sample.

E. CONCLUSIONS

1. It is possible to derive projections of the national crash experience from CARDfile.

2. Corrections to the national projection may be made on the basis of variation of CARDfile population characteristics from the nation.

3. Disagreement in accident subcategories with a small proportion of total accidents are much more likely to be due to sampling fluctuations rather than real differences in the true value. Such departures are more likely due to sample fluctuations of the small NASS sample.

REFERENCES

Edwards, M. A database for crash avoidance research. SAE, 1987???

Salvatore, S.; Mengert, P. and Walter, R. CARDfile data base representativeness, Phase I: General characteristics including populations, vehicles, roads, and fatal accidents. Transportation System Center, 1987.

ADDENDUM

CARDfile Variance Estimates

There is no very good way to get a standard error estimate for any CARDfile derived population estimate. The standard error derived solely on the basis of sample size consideration is grossly misleading. For example (See Table 1) 14.78% of CARDfile accidents are in the the "dark/lighted" category. If we took the standard sample size based estimate $\sqrt{pq/n}$ we would get .1478(.8522)/29000=.00066, i.e., 6 one hundredths of 1%. But the dark/lighted percent from state to state (see Appendix Table 1) varies from 8.60 (Texas) to 26.3 (Maryland).

In carefully designed experiments standard errors are often developed on the basis of which sampling units are included with certainty and which by chance. Here we do not have a designed experiment and individual states are in the system partly by design (i.e., to be representative) and partly by chance (i.e., because they were available). If we assume that the set of states is not very biased in its selection (Phase I gave some support for this assumption) then an expected overestimate of the standard error could be derived by assuming that each state is in the sample by chance. The simplest estimate for the standard error of some average quantity would then be based on the standard deviation of the average quantity for each state. For example the percents dark/lighted for the six states are as follows:

Indiana	18.33%
Maryland	26.26%
Michigan	12.34%
Pennsylvania	21.65%
Texas	8.60%
Washington	19.45%

The estimated standard deviation, S, is 6.385 by the formula . This is converted to a standard error by dividing by N, i.e., 6 in this case. The result is =2.6%. This is probably an overestimate but it is certainly of more usefulness than the 6/100 of 1% error obtained by the sample size based formula pq/n which clearly does not apply.