UM-HSRI-80-38

STATISTICAL ANALYSIS OF THE NATIONAL CRASH SEVERITY STUDY DATA

Phyllis A. Gimotty
Kenneth L. Campbell
Thipatai Chirachavala
Oliver Carsten
James O'Day

The University of Michigan
HIGHWAY SAFETY RESEARCH INSTITUTE
Ann Arbor, Michigan 48109

FINAL REPORT
August 1980

Prepared under contract No. DOT-HS-8-01944
Contract Amount $446,060

National Highway Traffic Safety Administration
Department of Transportation
Washington, D.C. 20590

Prepared for the Department of Transportation, National Highway Traffic Safety Administration, under Contract No. DOT-HS-8-01944. This Document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or the use thereof.

Technical Report Documentation Page

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

4. Title and Subtitle	5. Report Date
Statistical Analysis of the National Crash Severity Study Data	June 1980
	6. Performing Organization Code
7. Author(s) Phyllis A. Gimotty, Kenneth L. Campbell, Thipatai Chirachavala, Oliver Carsten, James O'Day	8. Performing Organization Report No. UM-HSRI-80-38
9. Performing Organization Name and Address Highway Safety Research Institute The University of Michigan Ann Arbor, Michigan 48109	10. Work Unit No. 11. Contract or Grant No. DOT-HS-8-01944 13. Type of Report and Period Covered Final
12. Sponsoring Agency Name and Address National Highway Traffic Safety Administration Department of Transportation Washington, D.C. 20590	14. Sponsoring Agency Code

Abstract:

This is the Final Report on a two-year statistical analysis of the data collected in the National Crash Severity Study (NCSS). The analysis presented is primarily concerned with the relationship between occupant injury severity and the crash conditions. Statistical models were developed to relate the probability of a severe injury to independent variables such as Delta V (the instantaneous change in velocity of the vehicle) and Occupant Age. Models were developed separately for various subsets of front and side impacts.

The area of population statistics was also studied. Design effects and variances were estimated for the aggregate NCSS data. Subsamples of both fatal and non-fatal cases with missing data on key variables were examined. This work supplements the separate publications of NCSS Statistics for passenger cars and for light trucks and vans. These publications present descriptive statistics on the accidents, vehicles, occupants and their injuries for the aggregate NCSS data.

Various clinical studies which were reported separately, are summarized in this report. These studies address the existing literature, side impacts, lower extremity injuries, eye injuries, and neck injuries.

The application of the NCSS data to accident analysis models, which estimate the benefits of improved occupant protection systems, is also discussed.

17. Key Words	18. Distribution Statement
NCSS statistical analysis, logit model, population statistics, design effects, injury severity, collision severity, clinical studies, missing data	Documentation is available to the U.S. Public through the National Technical Information Service (NTIS) Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	490	

This page is reserved for the Metric Conversion Chart.

ACKNOWLEDGMENTS

A project of this scope and duration inevitably involves most of the people at the Institute. The authors are indebted for all of the large and small favors received. We hope that no major contributions have been inadvertently omitted in the acknowledgments which follow.

Mr. James O'Day served as project director for the entire program. Responsibility for writing the report was divided among the other investigators.

The clinical studies were carried out under the direction of Mr. James O'Day and Dr. Donald L. Huelke. Others who contributed to the clinical reviews were Dr. John W. Melvin, Dr. D. Hurley Robbins, Dr. John D. States, and Dr. Robert A. Mendelsohn.. The format and organization of the NCSS Statistics publications were originally developed by Mr. James O'Day and Dr. Richard J. Kaplan. Ms. Leda Ricci was responsible for the production of the final editions. Assisting in this effort were Dr. Kenneth L. Campbell, Mr. Oliver Carsten, Mr. Joseph Andary, Ms. Kathleen A. Jackson, and Mr. Brian Wolf.

The analysis activities were organized by Ms. Phyllis A. Gimotty and Dr. Kenneth L. Campbell. Ms. Gimotty was responsible for the work in the area of population statistics. Ms. Gimotty was also responsible for statistical methodology to the mechanistic model development. She was assisted by Mr. Oliver Carsten who was responsible for the final file-building activities and carried out the sampling error computations. Dr. Campbell was responsible for the formulation of the work with mechanistic models and the evaluation of accident analysis models. Mr. Thipatai Chirachavala performed all of the statistical analyses required in the development and evaluation of the mechanistic models. Graphics were produced by Mr. Brian Wolf, who also assisted on many of the analysis activities.

The authors would like to thank Mr. David W. Smith at the Statistical Research Laboratory for making available necessary software for the logit analysis used in the mechanistic modelling. In addition Mr. Smith provided statistical consulting on the logit analysis and contributed current research on measurement error in the logit model to this project.

The authors appreciate the comments made by Mr. James O'Day and Dr. Keith Smith during the writing of this report. They would especially like to acknowledge Dr. James Hedlund, Contract Technical Manager, for his involvement and constructive comments throughout this project.

Finally, the authors wish to thank Ms. Sharon Derry for the word processing necessary to the production of this report.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 SUMMARY

This is the Final Report on a two year statistical analysis of the data collected in the National Crash Severity Study. The analysis presented is primarily concerned with the relationship between occupant injury severity and the crash conditions. The goal of this work is to further the development of techniques to estimate the benefits of improved occupant protection systems. These techniques require a statistical description of the national accident experience, and statistical models relating the collision event to the subsequent injuries.

The National Crash Severity Study (NCSS) was a major traffic accident investigation study conducted by the National Center for Statistics and Analysis (NCSA) of the National Highway Traffic Safety Administration (NHTSA). Data collection was initiated January 1, 1977 and ended March 31, 1979. An important innovation initiated in this study was the use of a computerized accident reconstruction algorithm developed by R. McHenry[1]. The principal output of this program is an estimate of the instantaneous change in velocity of each vehicle during the impact phase of the collision, referred to as Delta V. The NCSS is a precursor to the National Accident Sampling System.

Accidents were investigated in seven geographic areas within the continental United States. These areas were not selected at random, but rather were chosen because the NCSA judged that high-quality accident investigation teams could be readily established in them. Within each area, a stratified sampling plan was used to select accidents involving passenger cars, light trucks, and vans which were severe enough to require at least one of the vehicles to be towed from the scene. Pedestrian accidents, and other accidents in which an eligible vehicle did not have to be towed away, were excluded from this study. For the selected accidents, a common set of detailed information on the accident, the vehicles, the occupants and their injuries were collected.

---

[1] R. R. McHenry and J. P. Lynch, CRASH2 Users Manual, DOT/HS 802-106, November 1976.

Over the 27 month data collection period, 11,386 accidents involving 14,805 towed passenger cars and 24,976 occupants were investigated.

A complete summary of the work carried out is presented in this section. An overview is provided in Section 1.1, including a list of the resulting publications. Section 1.2 summarizes the results. Implications for future work, in particular the NASS program, are presented in Section 1.3. Finally, the organization of the remainder of this report is described in the last subsection.

1.1 Overview

The work carried out is organized into the five basic tasks listed below:

1. Develop statistical models relating the type and severity of impact to the probability of injury.

2. Develop population statistics from the NCSS data.

3. Produce a booklet of NCSS statistics for general use.

4. Perform a clinical analysis of selected NCSS cases to enhance current understanding of the occurrence of specific injuries and the associated injury mechanisms.

5. Review and evaluate existing accident analysis models.

The relationship of these basic tasks is illustrated in Figure 1.1.

Tasks 1 and 2 above were the major analysis tasks, and are shown in the top-center of the figure. Issues addressed in Task 2 included weighted analysis and the effect of the sample design on sampling errors, missing data, and the development of national projections from the NCSS data. A related task, Task 3, produced publications presenting descriptive statistics on the accident experience for the aggregate of the seven NCSS areas. The titles of these publications are:

NCSS Statistics: Passenger Cars, June 1980. Report No. UM-HSRI-80-36.

NCSS Statistics: Light Trucks and Vans, June 1980. Report No. UM-HSRI-80-37.

Statistical models relating the crash conditions to the probability of injury, Task 1, are referred to as "mechanistic models" in this

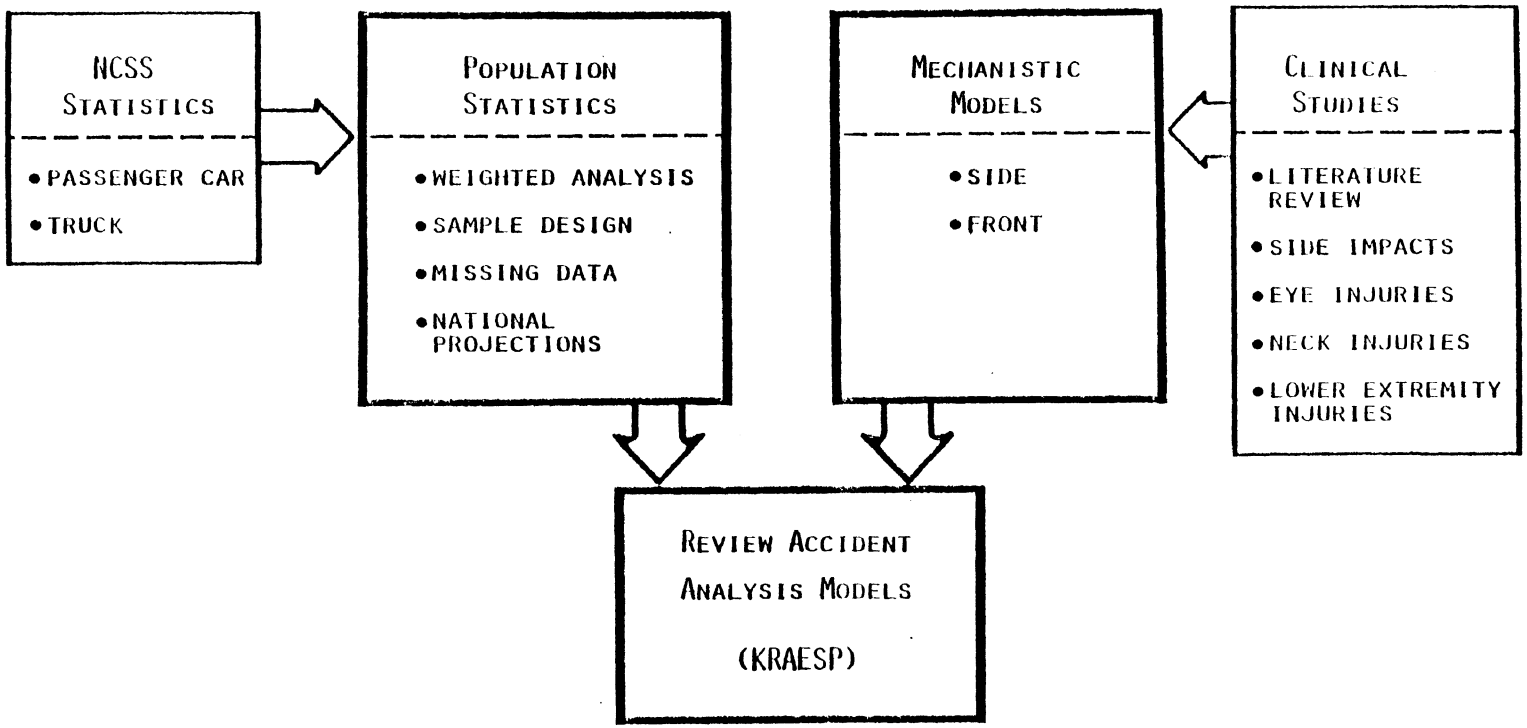

FIGURE 1.1 OVERVIEW OF PROJECT TASKS

report. This model development was carried out separately for occupants of vehicles with front as opposed to side damage. Each of these subsets was further partitioned depending on the occupant's seat location and a more detailed location of the damage. The major independent variable is the collision severity as estimated by Delta V, the instantaneous change in velocity of the vehicle.

The development of the mechanistic models relating the probability of injury to the crash conditions was approached from a deterministic point of view. The goal of these models is to reflect the relationships defined by the physical principles which govern the dynamic motion of the vehicle and its occupants during the impact, and the resulting injury mechanisms.

The fourth task addresses clinical studies. The importance of this task comes primarily from the need to provide a sound physical basis for the statistical development of the mechanistic models. Specific subsets of the NCSS cases were reviewed by a team of experts in the areas of biomechanics, dynamic testing, computer simulation of human surrogates, anatomy, and accident data collection and analysis to study the incidence of specific types of injuries and the relevant injury mechanisms. For example, side impacts were reviewed and compared with available side impact laboratory tests to determine if the actual injuries were comparable with test results. The objective of this review was to ensure that the statistical development of mechanistic models paralleled existing knowledge of the injury mechanisms and relevant variables.

The clinical studies resulted in several publications which are listed below.

Anatomy, Injury Frequency, Biomechanics, and Human Tolerances, NCSS Project Literature Review, May 1979. Report No. UM-HSRI-79-33.

Anatomy, Injury Frequency, Biomechanics, and Human Tolerances, February 1980. SAE Paper No. 800098.

Analysis of NCSS Side Impact Cases, August 1979. Report NO. UM-HSRI-79-50.

Side Impacts: A Comparison of Laboratory Experiments and NCSS Crashes, February 1980. SAE Paper No. 800176.

Lower Extremity Injuries in Automobile Crashes (An Analysis of NCSS Data), January 1980. Report No. UM-HSRI-80-10.

Ocular Injuries in Automobile Crashes, May 1980. Report No. UM-HSRI-80-22.

Cervical Injuries in Automobile Crashes, May 1980. Report No. UM-HSRI-80-40.

A major goal of the NCSS was to further the development of techniques to estimate the benefits of improved occupant protection systems. Current programs which carry out this estimation are called "accident analysis models." These models basically employ a statistical description of the current accident experience, and statistical models relating the probability of injury to the crash conditions, particularly the collision severity. The last task listed, Task 5, is a review and evaluation of current accident analysis models. This task is shown in Figure 1.1 as basically an application of the findings of Tasks 1-4.

1.2 Summary Results

This subsection summarizes the results obtained for each of the basic tasks. An important aspect of this work was the problem formulation and identification of appropriate analytical techniques. The reader is referred to Section 2 for a discussion of the conceptual approach employed.

1.2.1 Mechanistic Models. The model used in this analysis is the logit model. The logit model is a frequently used statistical model well suited to the analysis of frequency data. It uses the probability of only a non-severe injury (two or less on the Abbreviated Injury Scale[2]) as its dependent variable. Independent variables were both continuous and categorical, and included factors such as collision severity (Delta V) and Occupant Age. The critical aspect of the development of these models is the model evaluation. Both the Likelihood Ratio Statistic and the Goodness of Fit were examined. In particular, the predicted probability of injury was compared to the actual injury for each case, thus identifying "correct" and "incorrect"

---

[2]The Abbreviated Injury Scale, (1976 Revision). American Association for Automotive Medicine, Morton Grove, Illinois.

predictions. The mispredictions so identified were then examined in a manner similar to the residuals obtained in a regression analysis. Potential variables for addition to the models were reviewed in terms of their relationship to the mispredictions from the current model. Distributions of the predicted probability for the severe and non-severe injuries were also examined. This careful model evaluation procedure not only guided the model development, but also provided insight as to the limitations of the resulting models and the variables involved.

The NCSS data also provided an opportunity to evaluate the stability over time of the relationships modelled. Initial model development work was carried out with data from the first fifteen months of NCSS (referred to as Phase 1 for our analysis purposes). Subsequently, data for the last twelve months was received (Phase 2). The Phase 1 models were applied to the Phase 2 data and their predictive capability determined.

An important finding of this analysis is that the relationships modelled appeared stable over time, as evaluated by comparing results for the Phase 1 and Phase 2 data. The predictive capability of the Phase 1 models did not vary markedly when they were applied to the Phase 2 data. Furthermore, the predictive capability of comparable Phase 2 models was not appreciably different either. Of course, Phase 2 was just the continuation of Phase 1. There were no major perturbations such as an energy crisis or the introduction of new occupant protection systems. Nonetheless, this is the first data collection program of sufficient depth and scope to allow this kind of evaluation, and the results are encouraging. Subsequent statistical tests indicated no loss of predictive capability in the combined data set. The final models, which are presented in this summary, were all based on the combined data set, which comprises the entire 27 months of NCSS data collection.

The basic approach in the development of mechanistic models was to form subsets of the data which were expected to be more homogeneous with regard to injury mechanisms. The variables used to define subsets basically separated different collision types. Frontal-damage vehicles were modelled separately from side-impacted vehicles. In the frontal-damage group, vehicles striking other vehicles (two-vehicle accidents)

were separated from vehicles which struck other objects (single-vehicle accidents). In addition, center front impacts were distinguished from off-center impacts, as were drivers and right-front passengers. All side-impacted vehicles which were modelled were struck by other vehicles. Side-impacts to the passenger compartment were separated from non-passenger compartment impacts, as were occupants on the same side as the impact (near-side), and those on the far-side. Other collision types were not modelled.

Final models were developed for two subsets of side-impacted vehicles and four subsets of frontal-damaged vehicles. These subsets are listed below.

Side Impacts

1. Impacts involving damage to the passenger compartment and occupants on the same side of the vehicle as the impact (Near PCD).
2. All far-side occupants (seated opposite the impacted side) and near-side occupants for impacts not damaging the passenger compartment (Far Occ + Near NPCD).

Front Impacts

1. Single-vehicle accidents
2. Two-vehicle accidents with center impacts (CIA-2VEH)
3. Two-vehicle accidents with off-center impacts, drivers only (OID-2VEH)
4. Two-vehicle accidents with off-center impacts, right front passengers (OIP-2VEH)

The goodness of fit results for the two-variable models (Delta V and Occupant Age) are shown in Table 1.1 for the two side-impact subsets, and in Table 1.2 for the four front-impact subsets. Delta V is the dominant variable in these models. Logistic curves are shown for the first side-impact subset, Near PCD (near-side occupants of side impact to the passenger compartment) in Figure 1.2, and for subset 3 of the front impacts, OID-2Veh (drivers only, two-vehicle off-center impacts) in Figure 1.3. Confidence intervals are also shown on these figures. For convenience, the probability of severe injury is plotted. The effect of Occupant Age is much less than that of Delta V. The effect of Occupant Age is illustrated in Figure 1.4 for the first side-impact subset, Near PCD.

TABLE 1.1

Goodness of Fit

Severity = F(Lateral Delta V, Age)

Phases 1 and 2 - Side Impacts

Subset	Sample Size			Percent Correct Prediction	
	Non-Severe	Severe	Overall	Non-Severe	Severe
1. Near PCD	434	216	75.2	91.9	41.7
2. Far Occ + Near NPCD	1162	131	91.7	98.6	30.5

TABLE 1.2

Goodness of Fit

Severity = F(Delta V, Age)

Phases 1 and 2 - Front Impacts

Subset	Sample Size			Percent Correct Prediction	
	Non-Severe	Severe	Overall	Non-Severe	Severe
1. Single-Vehicle	1049	331	79.3	95.7	27.2
2. CIA-2VEH	1256	238	88.1	96.9	41.6
3. OID-2VEH	1412	176	92.4	98.4	44.3
4. OIP-2VEH	566	90	88.3	98.2	25.6

These tables reveal the major problem encountered in the early models; that is the relatively poor prediction of the occurrence of severe injuries. Although the overall percentage of correct predictions is about 90%, severe injuries are predicted correctly in only about 40% of the cases. Histograms showing the predicted probability of the injury being non-severe are shown in Figure 1.5 separately for those cases in which the actual injury was, in fact, "non-severe," and for

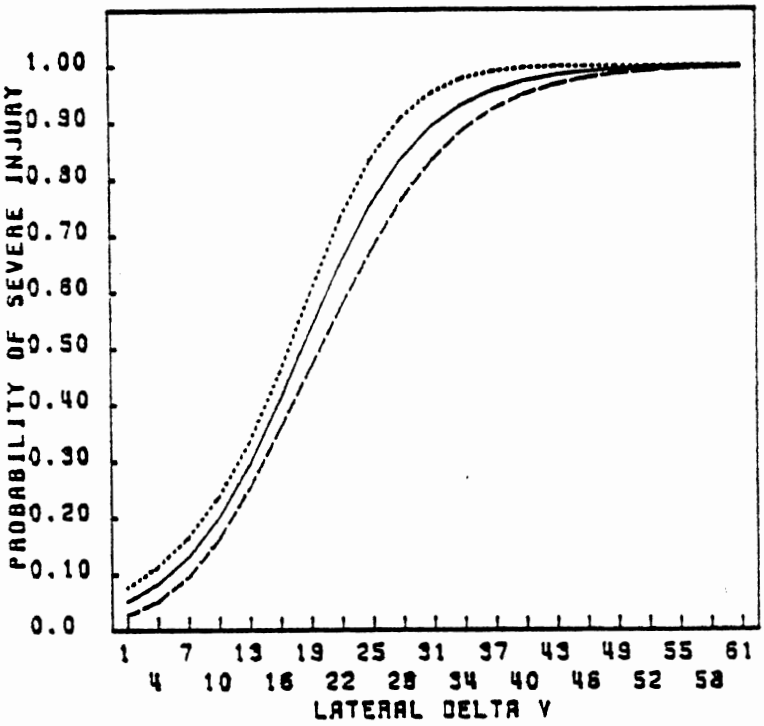

FIGURE 1.2 Logistic Curves of Two-Variable Models (Lateral Delta V, Age) For Near PCD Phases 1 and 2 - Side Impacts

those cases in which the actual injury was coded "severe." These histograms illustrate the information summarized in Table 1.1 (Near PCD).

A predicted probability greater than 0.5 was interpreted as a "correct" prediction if the actual injury was "non-severe," and "incorrect" if less than 0.5. Conversely, for the case where the actual injury was "severe," a predicted probability less than 0.5 was interpreted as a "correct" prediction. As shown in Figure 1.5, the predicted probabilities of non-severe injury are clustered nicely near 1.0 for the cases in which the actual injury was "non-severe." For the severe injury cases one would like to see predicted probabilities close to 0.0. As was shown in Table 1.1, only about 42% of the cases with severe injury were predicted correctly (predicted probability less than 0.5). The major thrust in the model development was improvement in the capability to predict the severe injuries. These, after all, are the cases one ultimately hopes to modify with improved occupant protection systems.

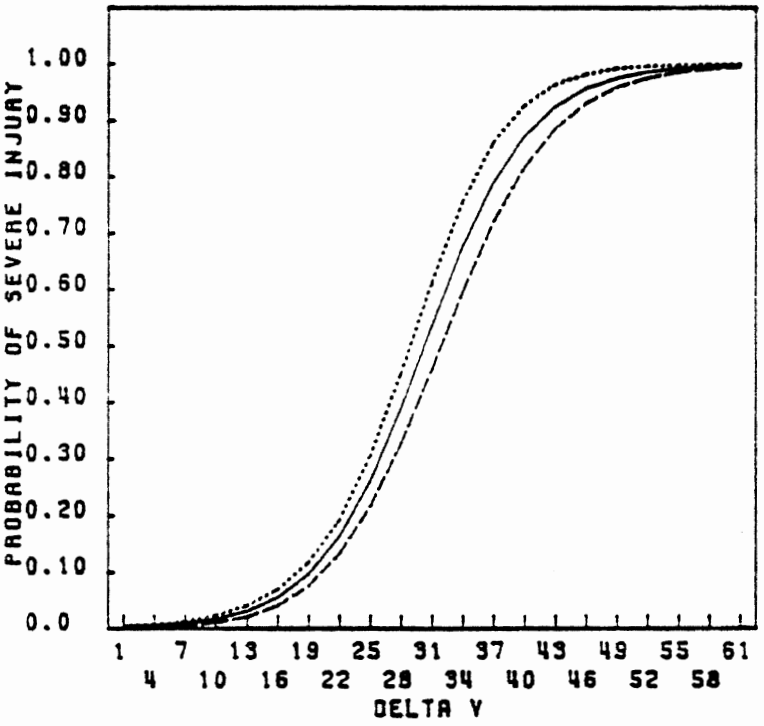

FIGURE 1.3 Logistic Curves of Two-Variable Models (Delta V and Age) For OID-2VEH Phase 1 and 2 - Front Impacts

Body Region (of the injury) and Injury Type were found to be highly correlated with the misprediction of severe injuries. Careful incorporation of these variables, particularly Body Region, improved the prediction of severe injuries substantially. The final model for the Near PCD subset predicted 67% of the severe injuries correctly and 89% of the non-severe injuries correctly, for an overall percent correct prediction of 81%. The improvement in predictive capability is apparent in the histograms of predicted probability shown in Figure 1.6, (again for the Near PCD subset). The effects of the various body region dummy variables employed are illustrated in Figure 1.7. Notice that the three dummy variables primarily involving extremities have approximately the same effect. The "other body region" group is primarily missing data on body region. This information is most likely to be missing in minor injury cases. Hence, the lower probability of severe injury.

The use of these injury variables is problematic. The Abbreviated Injury Scale is such that many injury types can only be assigned to one, or possibly two, AIS levels. In turn, particular injury types are

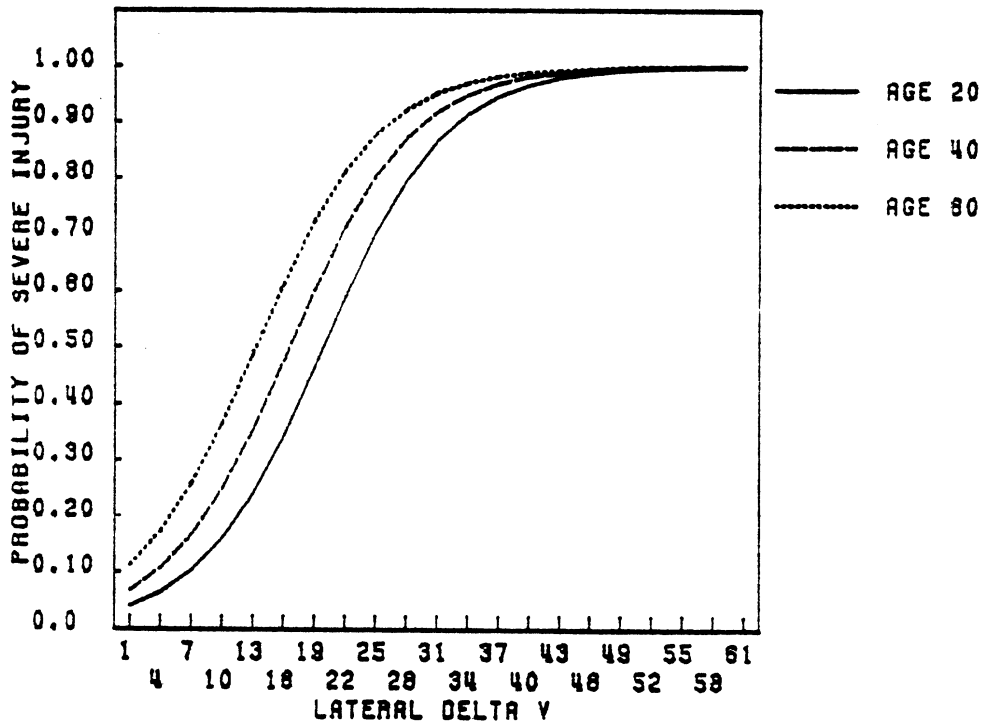

FIGURE 1.4 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Near PCD Phases 1 and 2 - Side Impacts

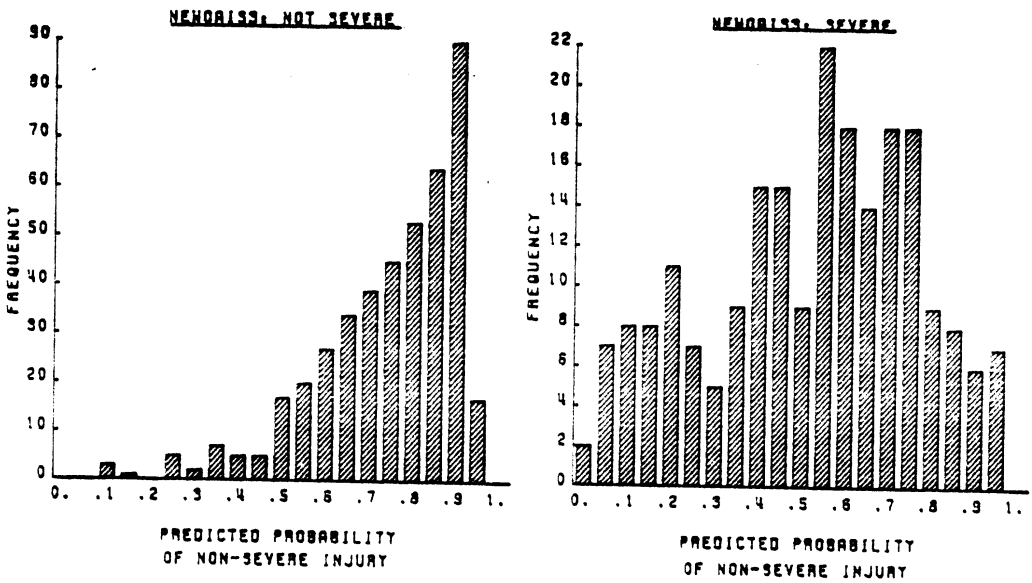

FIGURE 1.5 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For Near PCD Phases 1 and 2 - Side Impacts

associated with particular body regions; concussion occurs in the head, and fractures tend to occur in the extremities. Consequently, specifying the body region and/or injury type comes close in some

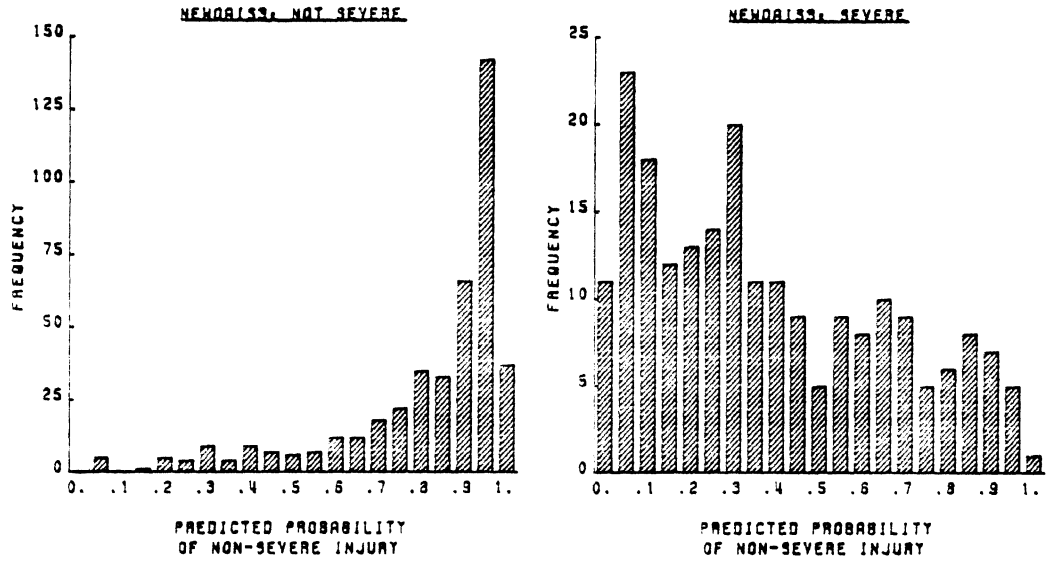

FIGURE 1.6 Histograms of $\hat{p}_i$ of Four-Variable Model (Lateral Delta V, Age, Object Contacted, Body Region) For Near PCD Phases 1 and 2 - Side Impacts

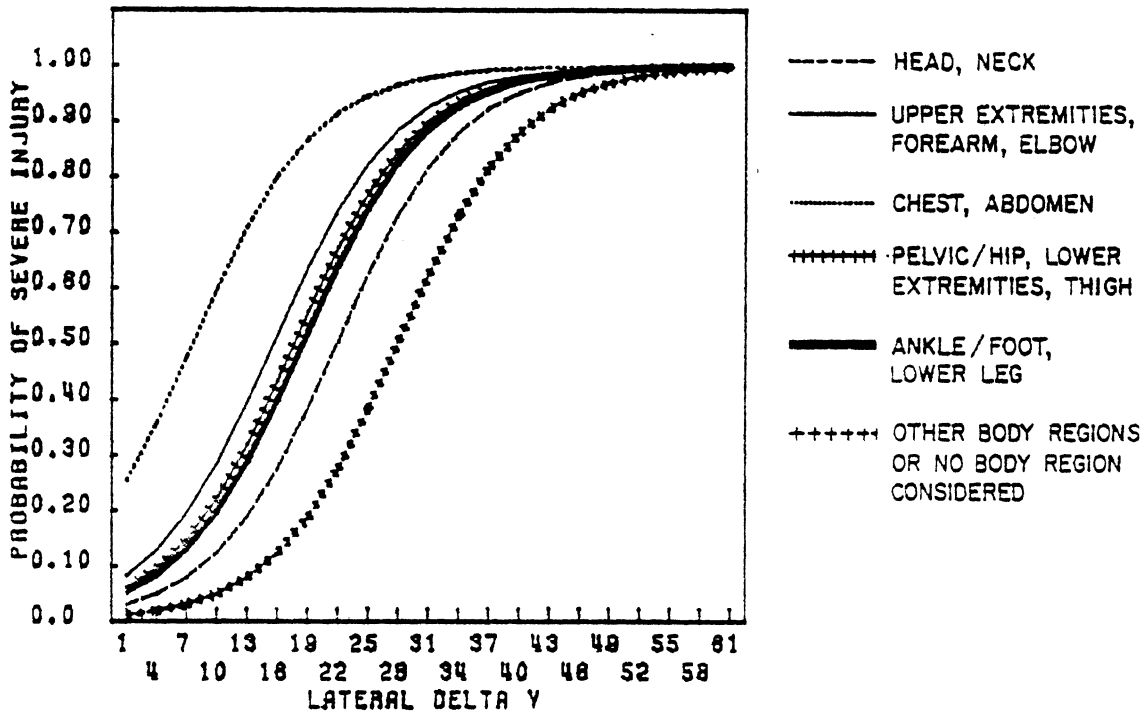

FIGURE 1.7 The Effect of Five Levels of Body Region of Four-Variable Model (Lateral Delta V, Age, Object Contacted, Body Region) For Near PCD Phases 1 and 2 - Side Impacts

instances to specifying the AIS level. From a practical standpoint, inclusion of body region or injury type in the model is not useful without knowing the factors which determine which body region is injured, or what the type of injury will be.

Other aspects of this problem seem to be inherent characteristics of the AIS scale. Mispredicted cases were usually relatively low to moderate collision severity (Delta V) impacts which resulted in severe injuries (OAIS 3+). For example, a substantial group of mispredictions in the front subsets were ankle dislocations. These receive an AIS 3 because of the joint involvement. Many factors were assimilated in developing the AIS scale: threat-to-life, treatment period, probability of permanent impairment, etc. Not all these factors are directly related to the collision forces, especially when comparisions involve different body regions and/or injury types.

In the context of the above discussion, the dummy variables corresponding to various categories of the body region variable were found to substantially improve the predictive capability of the models. This finding does not produce a satisfactory final model, since Body Region is basically a response variable. This finding is important because it clearly identifies the limitations of the current models, and suggests what refinements will be necessary to improve these models. This topic is pursued in Section 1.3, Implications for Future Work.

1.2.2 Population Statistics. One of the primary objectives of the National Crash Severity Study was to provide national estimates of totals and distributions of accident statistics for descriptive purposes and, more specifically for input into the accident analysis models. NCSS is a purposive sample of seven areas in the United States, chosen to represent the 1970 U.S. population, within which accidents were chosen using a stratified cluster sample. Since the seven areas were not chosen randomly there is no probability-based estimate available for national accident statistics. Adjustments to the NCSS data are necessary to make the NCSS statistics reflective of the national accident experience. The sample design will allow representative estimates for the aggregate of the seven areas.

Elements in the sample design and data collection procedure can have an effect on the reliability of accident statistics estimated. A good estimate for the aggregate is only possible if complete data are collected within each site. Missing data are of two types. If accidents are missed in the sampling process, there will be an

undercoverage of the population of interest. Incomplete data on some accidents is another potential source of missing data. Both of these types of missing data may affect the magnitude of accident statistics from NCSS.

One obstacle in the analysis of the NCSS data is the amount of missing data on crash severity and injury severity. A subsample of cases with incomplete data was obtained to evaluate possible biases introduced by the missing data. Fatal occupants were also compared with external sources to assess the magnitude of undercoverage in the fatal occupant population.

Fatal occupants in NCSS were compared with fatal occupants in FARS and state police files. A census of all fatalities was specified in the sample design and when matched with FARS it is possible that as many as 20% of the fatalities reported in FARS were not investigated in NCSS. This would result in an almost negligible effect on aggregate accident statistics if non-fatal accidents are not similarly under-reported, but would have a substantial effect when looking only at the fatal population of NCSS.

Missing data for fatal and non-fatal occupants were investigated separately. For the fatal occupants with incomplete data there was a higher proportion of occupants with OAIS coded maximum injury and an increase in the proportion of vehicles with fatal occupants at higher categories of Delta V. For non-fatal occupants there was a substantial difference in the proportion of minor injuries. When this information was combined and the distribution of OAIS adjusted for missing data, there was a substantial change in the proportions of no injury and minor injury for the NCSS aggregate data.

One of the main products of this task was the production of two publications of NCSS statistics. These publications are listed in Section 1.1 The publications describe police-reported accidents involving towed vehicles for the aggregate of the seven areas. In each publication tables are generally presented in two complementary forms. One page provides a frequency distribution of the factor under consideration; the opposing page shows the corresponding injury rates. Extensive graphical displays of the data are presented with the tables.

These graphical methods include histograms, bar graphs, pie graphs, line graphs and three dimensional plots.

In the publications of NCSS statistics, for completeness, missing data counts have not been excluded in the calculation of distributions. The focus of the missing data analysis was on differences between vehicles or occupants with incomplete data and those for which complete information was available. If no difference were apparent then the distributions for NCSS statistics ignoring missing data would be judged an appropriate description of the total population.

Even though sampling errors were not included in these publications, sampling errors were calculated for selected statistics. This investigation was done in order to assess the magnitude of the effect of the cluster design on the variability of various NCSS statistics. There were some statistics where the effect of the design on the variance was important. For most accident statistics the variances were two to three times larger than the variance calculated under the assumption of simple random sampling. Some statistics associated with injury severity actually had smaller variances then the simple random sampling variance. In general, variances calculated based on simple random sampling will underestimate the variability associated with the NCSS statistics.

Use of the NCSS data to produce accident statistics that are nationally representative involved the development of a procedure to adjust NCSS statistics. This procedure uses the NCSS aggregate accident statistics and demographic variables available for all areas in the United States to produce a national projection. This national projection method used the relationship observed with the NCSS statistics and the demographic variables to predict accident statistics for the unobserved areas of the country. These predicted values are then combined with the NCSS statistics to form the national projection. This method assumes that the relationship observed between the NCSS statistics and the demographic variable under consideration is the same relationship that exists in the unobserved areas.

The method to generate national projections was used for bivariate distributions as well as simple statistics. Comparison of this method

with a commonly used ad hoc procedure indicated these ad hoc estimates may underestimate the national level (or the national projection may overestimate). The method was evaluated to see how sensitive the national projection was to model choices. Empirically this procedure works well as long as the correlation between the accident statistics and demographic variables is high. This method does not produce stable projections when looking at events with a low probability of occurrence or with variables that have large amounts of missing data.

1.2.3 Clinical Analysis. This work complements the statistical development of mechanistic models. Identification of the injury mechanisms responsible for particular types of injuries guides the selection of variables and interpretation of outliers. As indicated in the list of publications presented in the Overview, Section 1.1, the clinical analysis was initiated with a review of the recent biomechanics and automotive injury literature. In general, a substantial gap was found between laboratory experiments and past medical studies of traffic accident injuries. The initial topic area selected for clinical analysis was the comparison of the NCSS side impact experience with available laboratory side impact simulations. Further clinical analyses addressed three specific body regions: the eye, the lower extremities, and the neck. The results of this work are briefly summarized in the following paragraphs.

The hard-copy file containing original coding forms and photographs were obtained for NCSS cases involving side impact. These cases were studied by the clinical team to compare the results of actual crashes with the existing laboratory test data.

Of the approximately 90 cases studied, 51 were judged comparable to the laboratory tests, which generally simulate a 90^{0} impact angle using an impact sled. The remaining cases usually involved cars struck at a point remote from the passenger compartment. The resulting rotation often appeared to influence the occupant's trajectory. Injuries for the 51 cases were tabulated by collision severity (Delta V). This result was found to be comparable to observed laboratory test results at slightly higher Delta V levels.

The second clinical study focused on severe injuries (AIS 3 or greater) to the lower extremities (pelvis, thigh, knee, leg, and ankle/foot). The lower extremities are the second most frequent body region receiving injuries at the AIS 3 or 4 level. The medical consequences of these injuries may be extreme, including prolonged immobilization, long recovery periods, and the potential for some degree of permanent impairment.

The more severe lower extremity injuries are most often sustained by unrestrained occupants impacting objects in front of them with the lower instrument panel being the most frequent contact point. Fractures are the most common type of injury in this study group.

Direct impact loading to any area of the lower extremities can cause injuries in that body region. However, it was also found that force transmission through bone to other lower extremity areas can cause fractures and/or dislocations remote from the impact site. Compression or twisting forces, especially at the ankle area, are believed to be the main cause of severe injuries to the ankle/foot region.

A review of NCSS cases involving eye injury indicated that the incidence of injuries to the eye is very low. There were 45 occupants with a reported eye injury out of a weighted total of 62,026 passenger car occupants in collisions severe enough to require the cars to be towed. The injury actually involved the eyeball in only 14 of the 45 cases. The low incidence of eye injury in the United States may be due to the use of "high penetration resistant" windshields. Tempered windshields commonly used in Europe have been shown to be highly related to ocular injuries[3]. About 50% of the eye injuries studied in the NCSS data were still caused by glass.

The NCSS data have also provided, for the first time, information on the frequency and severity of cervical injuries in traffic accidents. Approximately 0.3% (or one in 300) occupants of towed passenger cars received a severe cervical injury (AIS 3-6). However, among ejected

---

[3] M. Mackay, "Incidence of Trauma to the Eyes of Car Occupants." Trans. Opthal. Soc. of the United Kingdom, Vol. XCV, Part II, pp 311-314, 1975.

occupants, 7.2% (or one in 14) received serious cervical injuries. The clinical study revealed that the neck is rarely fractured or dislocated by direct impact to the cervical area. These injuries are usually the result of forces transmitted through the cervical spine as a result of head contact with the windshield, or other interior surface. The anterior neck structures, however, are almost always injured by direct blunt impacts.

1.3 Implications for Future Work

Using the NCSS data, this project integrated research in three areas: modelling of injury severity, estimation of statistics that describe the national population, and evaluation of current accident analysis models. The intent of this subsection is to review crucial areas identified in this research effort. In each of the following subsections tasks requiring research are described and our recommendations for further research are presented.

1.3.1 Modelling of Injury Severity. All models developed for predicting injury severity used a dichotomized version of the OAIS variable due to the amount of missing data on OAIS. This, in fact, limits the direct applicability of these models to the accident analysis models since these models require statistical models to predict the entire OAIS distribution. The most important observation obtained from the modelling of injury severity was that there were severe injuries that resulted from accidents with a low crash severity. Further investigation of this phenomena could be concentrated in two areas, modification of variables and development of more sophisticated models.

In upgrading the variables to be used in the modelling effort, the most important point is that every attempt to minimize the amount of missing data must be made. As indicated above there is evidence that injury severity does not strictly increase with crash severity. Variables other than Delta V, the only measure of crash severity in the NCSS data, may prove helpful in developing models consistent with the OAIS scale. Alternatively, additional work could be done with the OAIS scale to make it a multi-dimensional injury severity variable. This new variable may exhibit the expected association between injury and crash severity.

Further investigation into a more general class of models is a possible future activity that does not appear very promising for modelling a dichotomous injury severity model. Models to predict the entire OAIS distribution could eliminate the problem of predicted non-severe injuries at low crash severities when in fact the injury was severe. The most promising direction is to incorporate body region into the initial partitioning of the occupants into similar accident experiences when modelling injury severity.

Based on our experience with the NCSS modelling effort we would recommend further research in the following areas:

1. Model adjustments to incorporate missing data.
2. Definition of additional measures of crash severity.
3. Model development on partitions of the data that include collision type, seat position and body region.
4. Models developed to predict the OAIS distributions.

1.3.2 Estimation of the NCSS Statistics. NCSS statistics were developed to serve two purposes. These statistics, or a modification of them, were to be evaluated for use as basic input to the accident analysis models. The modification of these NCSS statistics to reflect the national accident experience was a primary task in this project. The other purpose of these statistics was to provide the highway safety research community a collection of accident statistics obtained in NCSS.

Each of these tasks involve analyzing the NCSS statistics to develop estimates reflective of the national accident experience. Models were developed using the NCSS statistics (as opposed to the raw data) to adjust the these statistics. The production of publications describing NCSS statistics pointed out problems that needed attention. The sampling error associated with all estimates from NCSS is necessary to evaluate the variability of the estimate. Missing data was substantial for some variables collected in NCSS and ignoring the missing data can distort the true distribution.

There are various recommendations we can make based on our experience with the NCSS statistics:

1. An imputation procedure be developed to adjust for missing data.

2. Design effects be periodically evaluated to determine whether they will provide a good summary for the effect of the design on the sample variance.

3. Investigate statistical methodology for population statistics and determine their relevance for statistical analysis in causation, crashworthiness, time series, and accident analysis models.

1.3.3 Accident Analysis Models. Two major points arise from this brief review of accident analysis models in general. They are:

1. National estimates of the accident experience are required.

2. The estimation procedures will be valid only to the degree that they reflect the actual physical principles and mechanisms which govern the events being simulated.

National estimates are required, since it is the national accident experience which is being projected. The important point here is that statistically based national estimates (which will eventually be available from NASS) carry with them estimates of their variance. If this information were carried through the simulation process, one would be in a much better position to evaluate the variability of the resulting projections.

The second point embodies the essence of what we have described as "mechanistic" models. For many applications, a statistical description of the current situation is completely adequate. The situation is much different, however, when one wishes to project the effect of proposed changes in the system. Statistical correlations present before the changes are introduced may be altered. Controlled experiments generally cannot be conducted in a social system. The alternative is to ground the statistical models in the physical principles and mechanisms which govern the event being simulated. This is the critical issue in the projection of the accident experience of the hypothetical vehicle population, and also in projecting the injury response of the proposed restraint systems.

The KRAESP[4] model was reviewed in more detail. Specific implications for the future use of this model follow.

In the crashworthiness area (the projection of injury response), the subsets used by the KRAESP model are generally comparable with those which evolved from our work. The important observation here, is that the prediction of severe injuries was correct only about 40% of the time unless body region was included in the model. The implication is that separate models should be developed for at least three or more generalized body regions. Since not all injuries are coded for the NCSS data, separate injury distributions for each body region may be somewhat underestimated.

Another important issue concerns the possibility of year-to-year changes, or trends, in the national accident experience. Currently, the KRAESP model uses a single set of accident statistics. Only the total number of accidents is adjusted to reflect estimated changes in total vehicle use. If vehicle populations are hypothesized with appreciably different distributions of car size, then it might also be reasonable to envision that the use of these vehicles may be different. Adjusting the distributions in the national accident experience to reflect year-to-vear changes is a subset of the larger problem of accident causation. The current adjustments to the total number of accidents from year-to-year represent a simplified model of accident causation. The implications of these assumptions need to be reviewed.

A final observation is that missing data will be a serious problem. Either Delta V or OAIS are missing on 60% of the file. Our modelling efforts only addressed the most promising front and side impact subsets. Alternative techniques will be needed where Delta V is not a suitable measure of collision severity. Statistical techniques such as those discussed in Section 4.1.4 or 7.5 will have to be employed to address the missing data problem.

---

[4] D.Redmond and K. Friedman, "Introduction to the Kinetic Research Accident Environment Simulation and Projection Model," Prepared under DOT Contract No. DOT-HS-9-02096, Kinetic Research Draft Report No. KRI-TR-041, January 1980.

1.4 Report Organization

The major tasks in this project are organized under the topical headings, "mechanistic models" and "population statistics." This material is presented in Sections 3 and 4 respectively. Development of mechanistic models was pursued separately for vehicles with front as opposed to side damage. Preliminary model development for each was carried out with a preliminary file containing data from the first fifteen months of the NCSS. Final model development was based on data from the entire twenty-seven month study. The subsections of Section 3 reflect these divisions.

The initial subsections of Section 4, Population Statistics, address the use of the sampling weights and the influence of the sample design on the variance of NCSS estimates. Various approaches to the problem of missing data are discussed in this section, as well as the development of national projections from the NCSS data.

A review and evaluation of accident analysis models is presented in Section 5. Section 6 summarizes the clinical studies which were conducted. Finally, implications for future work and the analysis of the NASS data are discussed in Section 7.

Appendices contain the NCSA algorithms for generating the "NEWOAIS" variables, the data structure for variance computations, and tables of estimated variances and design effects for selected statistics.

2 CONCEPTUAL APPROACH

This section describes the overall conceptual approach which underlies the work presented in this report. The uses of accident data in highway safety research may be broadly divided into analysis of accident causation and analysis of vehicle crashworthiness. This project is concerned with the crashworthiness problem, which may be defined as an evaluation of the ability of the vehicle to protect occupants from injury, given that a crash occurs. The ultimate objective is the evaluation of improved occupant protection systems. An intermediate goal, the goal of this project, is simply the development of basic analytical tools. The presentation of the conceptual approach begins with a restatement of the study objectives. These objectives, in turn, define the research areas to be addressed.

The primary objective of the NCSS study[5] was to assemble a data base to verify and/or refine procedures for estimating benefits of potential countermeasures in the crashworthiness area (vehicle structures and occupant protection). Current programs which carry out this estimation procedure are called "accident analysis models." These models, in turn, rely on a statistical description of the current national accident experience, and statistical models relating the collision event to the subsequent injuries.

The basic task statements are repeated here:

1. Develop statistical models relating the type and severity of impact to the probability of injury.

2. Develop population statistics from the NCSS data.

3. Produce a booklet of NCSS statistics for general use.

4. Perform a clinical analysis of selected NCSS cases to enhance current understanding of the occurrence of specific injuries and the associated injury mechanisms.

5. Review and evaluate existing accident analysis models.

---

[5] C.J. Kahane, R.A. Smith, and K.J. Tharpe, The National Crash Severity Study, Proceedings of the Sixth International Technical Conference on Experimental Safety Vehicles

These tasks have been interpreted in the context of the overall NCSS goal which is the evaluation of potential countermeasures in the crashworthiness area. Accident analysis models embody current efforts to bridge the gap between laboratory testing and the real-world accident experience. Laboratory testing of vehicle crashworthiness is necessarily limited because human test subjects cannot be used. For the most part, field accident data is the only source of information on the nature and consequences of severe impacts to humans. Of course, the accident experience is also of interest because it is the target population one is trying to modify.

The central function of an accident analysis model is the estimation of the resulting injuries for a prototype occupant protection system when implemented in the current collision environment. This estimation can only come from a physical model of the injury mechanism which relates the dynamic motion and forces of the impact to the resulting injuries. Current accident analysis models derive these physical models of the injury mechanism from existing accident data.

The primary objective of the statistical models listed in the first task above is to fill this need. The major hurdle in using accident data for this purpose stems from the fact that real accidents do not occur under the same controlled circumstances as laboratory impacts. From an analysis point of view, it should be recognized at the outset that accident data must be thought of as data from an unplanned experiment. Most of the variables in the models are there as control variables, rather than variables over which the vehicle designer has any influence. Variables of interest will sometimes be correlated with one another so that their effects are not readily distinguished. The most important point is that the relationships observed cannot prove a cause and effect relationship because the independent variables were not deliberately manipulated. Any inferences of a causal nature must be based on an understanding of the physical mechanisms governing the collision events. Such models are referred to as "mechanistic" models in this report. This deterministic view provides the basis for the development of these models.

Accident analysis models also seek to project the change in the overall accident experience due to the introduction of improved occupant protection systems to a portion of the total vehicles and occupants involved. In order to do this, a statistical description of the current accident experience is required. This description includes various totals, such as the total number of accidents, vehicles, occupants, injuries, fatalities, etc., as well as distributions for various descriptive variables such as collision type, direction of principal force, damaged area, collision severity, etc. Hence, the question of developing population statistics from the NCSS data is the second major task listed.

The final task brings the information gained in the preceding tasks to bear on existing accident analysis models. The objective here is to review and evaluate existing models, in particular the Kinetic Research Accident Environment Simulation and Projection (KRAESP) Program[6], in light of the findings of this study.

The organization of work carried out, and the organization of this report, parallel the tasks which have just been outlined. In initiating this work, the first step was the identification of suitable analytical methods. This determination is primarily derived from the overall problem formulation.

The remainder of this section describes the analytic approach that was taken. The central objective of this study is the development of statistical models relating the collision variables to the probability of injury, and the use of these models to project the effects of improved occupant protection systems to the national accident experience. Current computer algorithms to carry out this projection are called "accident analysis models." A conceptual overview of their function is presented in Section 2.1. This overview forms a background for the analytic approach presented in the remainder of this section.

---

[6]D. Redmond and K. Friedman, "Introduction to the Kinetic Research Accident Environment Simulation and Projection Model," Prepared under DOT Contract No. DOT-HS-9-02096, Kinetic Research Draft Report No. KRI-TR-041, January 1980.

Section 2.2 addresses the conceptual issues and available analytical techniques in relation to the development of population statistics from the NCSS data. A related issue involves the population described by the NCSS Statistics tabulations prepared for the aggregate of the NCSS sites, and the computation of variances for these results.

Section 2.3 describes the problem formulation and analytic techniques employed in the development of mechanistic models relating the collision variables to the probability of injury. As with the topic of national estimates, this work was pursued primarily in light of the application to accident analysis models. Hence, this approach follows the discussion of the objectives of accident analysis models presented in Section 2.1. This statistical development of models predicting the probability of injury is complemented by the clinical review of hard-copy NCSS cases. The objectives of the clinical reviews are summarized in Section 2.4.

The final subsection, 2.5, provides a brief overview of the NCSS data. In particular, the various data files used over the course of this work are described and the sampling strata are briefly reviewed.

2.1 Accident Analysis Models

The objective of these models is to estimate the reduction in deaths and injuries for future populations of vehicles with various improved occupant protection systems. Projections are computed for each calendar year for several future years (through 1990 in the KRAESP model). Each year, the vehicle population is revised to include new vehicles introduced and to drop older vehicles which are scrapped. The estimated benefits of the improved occupant projection systems are then presented as trends in the estimated total number of deaths and injuries over several calendar years as the composition of the total vehicle population includes more and more vehicles with improved occupant protection systems. The remainder of this discussion will focus on the estimation process for a single calendar year, since this same process is simply repeated for each successive year.

In general, these models must synthesize the current information in both the accident causation and the vehicle crashworthiness areas.

Assumptions in the area of accident causation are necessary when the number of vehicles (or their estimated annual mileage) in the future vehicle population is appreciably different from the current one. In this situation, the total number of accidents is adjusted accordingly. A more complete discussion of accident analysis models is presented in Section 5.

Focusing on the portions of accident analysis models dealing with the crashworthiness area, one finds the need for two basic types of information.

1. National estimates of the current accident experience, and
2. Mechanistic models relating the probability of injury to the various collision variables.

Each of these information requirements will be briefly described.

The current national accident experience, adjusted to reflect no use of available restraints, provides the baseline for projection of the benefits of improved occupant protection systems. Various adjusted population totals, such as the total number of accidents, deaths, and injuries, are needed. This description of the national accident experience must also be of sufficient detail to identify the individual subsets of the total accident experience for application of the mechanistic models. This description may take the form of a multivariate distribution of the various categories of the variables which define these subsets. Examples of these variables are:

1. Collision mode (single-vehicle, two-vehicle)
2. Vehicle damaged area (by clock direction)
3. Occupant seat position

In addition, distributions of closing speed are required for each collision mode and damaged area, and distributions of AIS are required for each five mile per hour increment of Delta V for each occupant seat position.

The subsets essentially define the mechanistic models. Within each subset, the probability of injury is presumed to be a function of only the collision severity as measured by Delta V. Consequently, the

"mechanistic model" is defined, for the baseline population of unrestrained occupants, by the distributions of AIS for each subset and for each 5 mph increment of Delta V as contained in the current national accident experience. A separate portion of the model estimates these same distributions for the prototype occupant protection systems.

Total numbers of injuries and deaths are estimated from these intermediate results. The projected vehicle population identifies which occupants of which vehicles will be protected by the improved systems. For these occupants, estimated distributions of AIS are substituted for the baseline (unrestrained) distributions. The estimated total numbers of injuries and deaths are simply obtained by summing the estimated numbers for each of the subsets.

In summary, the accident analysis models partition the overall accident experience and associate a particular distribution of AIS with each level of collision severity (5 mph increment of Delta V) within each subset. The performance characteristics of prototype occupant protection systems are quantified by associating revised distributions of AIS with a given collision severity level and subset. This analysis corresponds to the mechanistic models developed as part of this project. In each case, a cause and effect relationship between the probability of injury and the independent variables is assumed. The independent variables are the collision severity and the variables which define the subsets.

Two critical issues must be kept in mind when considering the use of these mechanistic models in the accident analysis models. The straight-forward issue is the evaluation of the predictive capability of the mechanistic models in the NCSS data file (the file used to develop the models). The more difficult issue is the suitability of the models for prediction of the injury experience beyond the NCSS file. This issue ultimately comes down to a subjective assessment of the degree to which the model reflects the physical mechanisms and principles which govern the collision event, and, in turn, the degree to which the injury mechanisms operating in the NCSS file will be appropriate to the accident experience of vehicles equipped with different occupant protection systems.

The formulation of the mechanistic model development is presented in Section 2.3, while Section 3 presents the results of this effort. The statistical development of mechanistic models is complemented by the clinical studies which attempted to identify injury mechanisms associated with particular types of injury. The objectives of the clinical studies are summarized in Section 2.4, and the results of this work are presented in Section 6. The following subsection describes the analytical approach followed in the area of population statistics.

2.2 Population Statistics

In order to assess the potential for injury reduction in the national accident population using the accident analysis models, a good description of the national accident experience is necessary. The best description of the accident population involves a probability based estimate of totals and distributions for vehicles and occupants involved in accidents. Necessary distributions for vehicles include model year, weight class, and Delta V. For occupants, distributions for seat location, collision mode and severity and damage area are necessary. In order to project injury reduction, national totals of vehicles and occupants are required. All of these descriptive statistics of the accident population and relationships between these population statistics will be referred to as "population models." This requirement of the accident analysis models led to the investigation of methods for obtaining estimates from the NCSS data of the national accident population.

As national probability samples become more and more prevalent it is becoming increasingly important to make a very clear distinction between different types of "national estimates." A national estimate is a number that is descriptive of some aspect of the national experience (e.g. the total number of accidents involving a towed vehicle). Estimates may be developed in more or less sophisticated ways, and, if they predict totals for the U.S., can be called national estimates. One possible estimation method combines available data (usually census-type information), subjective information, and judgment. An example of this method is to use gasoline tax receipts and estimated vehicle fuel

consumption rates to estimate the total number of miles driven by the U.S. population.

On the other hand, national estimates may be generated using data obtained from a national probability sample. Such data are obtained following the specifications of a sample design. The design usually includes information about the population sampled that helps to insure a well-distributed sample. The most valuable characteristic of a probability sample is that it is objective in its choice of sampled elements. Any bias that results in the estimate is due to chance. A national estimate generated using these procedures should be acknowledged. An estimate based on a national probability sample will carry along with it qualities that will increase it credibility. In addition all such probability-based national estimates will have a sampling error associated with them that can aid in assessing the reliability of the estimates.

The NCSS data provides probability-based estimates only for the seven areas that were chosen for inclusion in NCSS. These estimates can be aggregated to form statistics that describe the aggregate accident experience for the seven areas chosen. The seven areas in NCSS were chosen purposively, and therefore inference to the national population is not possible within the context of the sample design.

Methods to enable the NCSS data to be used to produce projections of the national accident experience were investigated. The technique developed uses relationships between NCSS statistics and demographic variables to predict accident statistics for those areas not investigated. These predictions are then combined with the NCSS data to form a national projection. As with all estimates of this type the national projection is biased. However, the context within which the national projection is developed provides a reasonable method for obtaining estimates for the national population.

Vehicles and occupants with incomplete data pose a serious problem in NCSS. For key variables there is a substantial amount of missing data. This not only affects the estimates made for the aggregate of the seven areas but also the national projections derived from these variables. If the amount of missing data is small, less than 5%, most

analysts would consider ignoring the missing data. With substantial amounts of missing data, ignoring it may cause biases in the resulting distributions. An alternative approach is to employ an adjustment procedure. All of these imputations and reweighting procedures assume that missing data is missing at random. This assumption is violated when certain types of accidents, vehicles or occupants are systematically excluded.

More information was needed in order to assess the alternative approaches to the incomplete data problem. Investigation into the biases introduced by the missing data was facilitated by a reevaluation of hard-copy case material for a subset of the cases with missing data. This approach provides information about possible bias introduced by ignoring missing data and is essential in choosing a missing data adjustment procedure.

To use the NCSS data as input to the accident analysis models, adjustment must be made for missing data. The missing data on key injury and crash severity variables is substantial enough to affect the NCSS distributions. These distributions, adjusted for missing data, can then be used to produce national projections. For this application, the projection method was modified to incorporate adjustments for missing data of various types to produce adjusted national projections.

There is no method available to assess how close any adjusted estimates come to the true national level. For a national projection without missing data, an estimate of its variability can be calculated. With missing data, sensitivity analyses will give more information about the variability inherent in the projection method and missing data adjustments.

The variability of these estimates is also important. The result of the accident analysis models is an assessment of potential injury reduction. If the estimated distributions and totals are highly variable then the result from the accident analysis models will be subject to, at the least, the same amount of variability. If measures of variability for statistics used as input into the model are available, then work can be done to assess the effect of this variability on the assessment of potential injury reduction.

A related but separate task in this project required the organization and publication of accident statistics from the NCSS data. These statistics are published for the aggregate of the seven NCSS sites. In these publications no attempt was made to adjust for missing data. Missing data was treated as a separate category and was not excluded in the calculation of the distributions. No sampling errors were included in these publications.

However, sampling errors for the NCSS statistics were evaluated. Methods of summarizing sampling errors for use in publications such as those produced for this task were examined. It is important to note that the NCSS design is a cluster design and hence the sampling errors are likely to be larger than those under a simple random sample design.

2.3 Mechanistic Models

Mechanistic models have been defined in the context of accident analysis models in Section 2.1. The objective is to develop statistical models which reflect the physical principles which govern collision events and injury mechanisms. These models provide a better understanding of the factors which influence the resulting injuries, and consequently lead to the development and evaluation of improved occupant protection systems. For this expectation to be realized, the statistical models must reflect the governing physical principles. In current practice, these considerations guide the selection of variables and the formulation of the model.

One immediately observes that the consequences of automobile accidents vary greatly. It is also clear that much of this variation arises from differences in the type and severity of collisions as well as the effects of other uncontrolled variables, rather than differences in vehicle crashworthiness. The underlying view is that there is a physical model for the entire collision process up to and including the actual injuries. This physical model provides a mathematical basis for the models to be developed. Unfortunately, the actual expression of such models is substantially beyond the current state-of-the-art. However, some information is available on the variables which are likely to be included and relationships between some of them. The starting point is to assume that the physical model and the associated

mathematical function exists. The dependent variable is injury severity. The first step is to identify the independent variables. It is convenient to group the independent variables as descriptors of the collision, the vehicle, and the occupant, as shown in Table 2.1.

TABLE 2.1

Model Overview

Dependent Variable	Independent Variables
Injury Severity	Collision
	Vehicle
	Occupant

This grouping may be related to a simplified block diagram of the collision event shown in Figure 2.1. The input is the collision itself: the vehicles and objects involved, their speeds, orientation, etc. For a particular vehicle, this collision is an input to the vehicle structure which ultimately transfers a deceleration pulse to the occupant. In general, the deceleration time-history experienced by the occupant is not the same as that of the vehicle center of mass; different occupants in the same vehicle may experience different pulses, and different parts of the occupant's body will experience different decelerations. The biomechanical characteristics of the occupant determine the injuries which result from the decelerations experienced. Laboratory simulations generally reproduce the collision input, control the vehicle parameters, and observe the deceleration of test dummies, or other human surrogates. Information on human tolerance to injury has generally been collected by observing the injuries sustained by test animals or cadavers under known deceleration levels.

Examples of independent variables in each of the three groups are shown in Table 2.2. Perhaps the most important control variables are those in the collision group. The collision variables are intended to

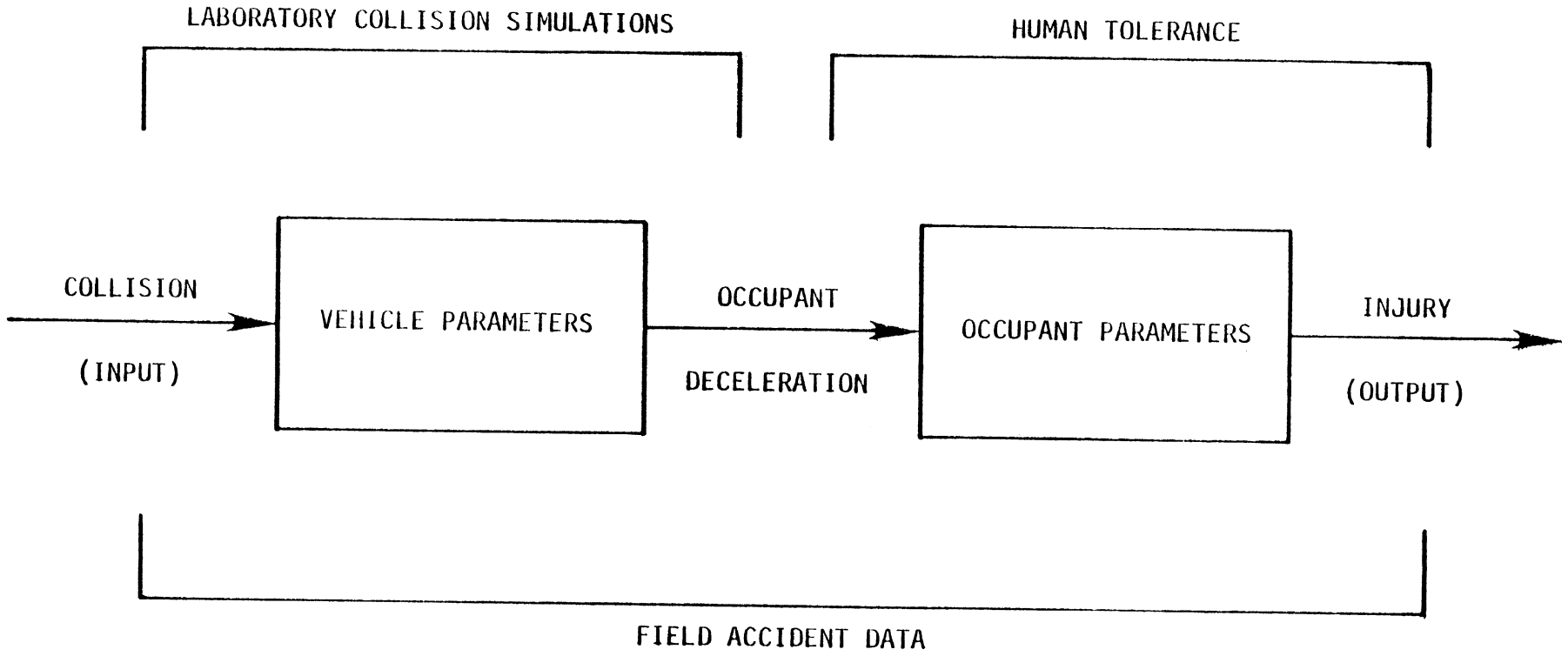

FIGURE 2.1 Simplified Block Diagram of the Collision Event

specify all aspects of the collision which influence the resulting injuries. They include the collision configuration, vehicle speeds, orientations, etc. In general, this group of variables includes the kinds of information an engineer would need to know in order to stage, or reproduce, the impact with test vehicles.

TABLE 2.2

Independent Variables

Collision Variables	Vehicle Variables	Occupant Variables
Severity	Mass	Age
Delta-V		
CDC Extent	Car Size	Sex
Rotation		
	Restraint	Height
Mode (Type)	(Ejection)	
G.A.D.		
Obj. Contacted		Occupant Location
P.D.O.F		
Rural/Urban		

The most important of the collision variables is the collision severity. With all of the other collision variables held constant, one would expect the potential for injury of the impact to increase with increasing levels of collision severity. Several alternative measures of collision severity are available. The primary measure is Delta V, the velocity change during impact. However, this variable has considerable (40%) missing data and is not defined for all collision types (rollover, for example). The CDC extent number is a possible alternative for these cases. Another alternative is the energy absorbed, which may also be expressed as an "equivalent barrier speed." This variable would have less missing data than Delta V because damage information on only the vehicle of interest is required. However, the effect of differing vehicle weights is not taken into account. While some comparison of the relative utility of these variables as predictors of injury severity is useful, in the end a single variable should be

selected, since each basically quantifies the same aspect of the collision.

There are, however, other potential variables which address what might be called second-order effects of collision severity. An example of this type of variable is the change in angular velocity. The influence of this variable is additive with Delta V. The magnitude of this added factor is proportional to the product of the angular velocity and the distance from the center of mass to the point of interest.

The remaining collision variables define collision types, or modes. These variables are necessary because the occupant trajectories, contact points, and injury mechanisms vary from one collision type to another. This variation is expected to substantially alter the relationship of injury severity to collision severity. The objective in defining collision types is to form subsets where occupant trajectories, contact points, etc. are sufficiently similar that a single model is expected to be adequate. The limiting factor is that sample size must be "adequate" within each collision type. Compromises must be made between the "homogeneity" and the size of the subsets. Potential subsets may be evaluated by comparing the relationship of collision severity and injury severity. For example, a scatterplot of the proportion of severe injuries may be examined.

The first level of subsetting separates frontal-damage vehicles from those with side damage. Within each of these groups, single-vehicle accidents are separated from two-vehicle accidents. Additional variables used to subset cases include the occupant's seat location and a more specific damage location (center, off-center, near-side, far-side).

The variables in the vehicle group are intended to define all the characteristics of the vehicle which influence the transmission of the vehicle impact to the occupant. Included are such factors as the vehicle's structural characteristics, occupant compartment space, restraint systems, etc.. All of the variables amenable to modification by the vehicle designer are in this group.

The variables in the occupant group include those factors which influence the occupant's injury thresholds. These include age, sex, pre-existing medical conditions, etc.

As already mentioned, the initial models look only at collision severity and injury severity as the selection of the variables defining the collision type subsets is refined. Next, variables from the vehicle and occupant groups were tried in an attempt to improve the predictive power of the model.

Additional considerations to keep in mind while developing these models are: the accuracy of measurement, or quantification, of the variables, especially Delta V, the adequacy of the functional form of the model, and the weighting, or distribution, of the cases across the various levels of the independent variables. Systematic errors in the estimation of Delta V are likely to be associated with car size and area of damage (that is, front, side, etc.) if the assumed structural stiffness is incorrect for some vehicles. A consistency check for this particular type of systematic error could be made by comparing the peak force computed by the CRASH2 program. The peak force should be approximately the same for the two vehicles. Large deviations may suggest problems with the assumed constants or input measurements for the Delta V computation.

The logit model has been selected as most appropriate for this work. The dependent variable, the probability of a non-severe injury (OAIS 2 or less), is dichotomous. This model describes the situation where the probability of injury varies systematically from case-to-case with other independent variables which have been observed. A complete description of this model is provided in Section 3.1.

The mechanistic model development was carried out in two phases which correspond to the preliminary data sets which were received. The initial work used a near-final version of the data collected during the first fifteen months of the NCSS study. Subsequently, data for the last twelve months were received. This situation provided an opportunity to "validate" the Phase 1 models.

Validation of the Phase 1 models was carried out to assess the apparent stability of the models over time. This is deemed to be a very important test. If stability cannot be demonstrated, the models would be of limited value. The NCSS study is the first data program of sufficient depth to be maintained over a long enough period of time to provide an opportunity for this kind of assessment. The test is carried out by applying the Phase 1 models to the Phase 2 data and determining their predictive capability. This result is compared to both the original Phase 1 results and the comparable model in which the coefficients are determined for the Phase 2 data.

Another consideration to be mentioned here is the weighting, or distribution of cases among the various levels of the independent variables. In general, the over-sampling of cases at higher injury levels shifts the distribution of cases towards higher severity levels. However, the objective of the mechanistic models is to approximate the theoretical relationships. The higher severity levels are of particular interest. Although an optimum design matrix is dependent on the specific model chosen, it generally is desirable to have the data evenly distributed among the various levels of the independent variables. Currently the raw data is heavily skewed to the lower severity cases. Weighting would aggravate this already uneven distribution. Therefore, the unweighted data are more appropriate for the development of mechanistic models. In fact, further subsampling of cases to achieve a more uniform distribution of cases across the levels of the independent variables was also tried.

A final consideration is the influence of substantial missing data. This subject is addressed in relation to population statistics in Section 4. No efforts were made to incorporate adjustments for missing data in the development of mechanistic models. If true physical relationships are reflected by the models, then missing data should not be a particular problem, unless specific regions of the domain of interest are excluded.

Section 3 contains a complete presentation of the mechanistic model development. Particular attention is devoted to the model evaluation procedures. As a consequence, this effort provides a complete

assessment of the problem of modelling the probability of injury. The final subsections which follow describe the objectives of the clinical work, and briefly describe the NCSS data files which were used.

2.4 Clinical Review

The clinical work complements the statistical development of mechanistic models. This work was initiated with a review of the recent biomechanics and automotive injury literature. The purpose was to determine the existing knowledge relative to injury type, body region, frequency, severity, and cause, and to identify gaps in the current knowledge that might be addressed with the NCSS data. In general, a substantial gap was encountered between laboratory experimentation and actual accident experience. Typically, the laboratory testing is precisely conducted with abundant data collection. In contrast, no controls are associated with the real accident experience, and virtually no information is available on the dynamics of the event. A major innovation of the NCSS study is the use of computerized accident reconstruction techniques which provide an estimate of the instantaneous velocity change during impact. This information materially aids the comparison of laboratory and actual accident experience. Selected side impact cases were chosen because of the current interest in this area of occupant protection. Other areas identified in the literature review were injuries to eye, lower extremities, and the neck. These later three areas were studied to identify the injury types and mechanisms associated with the selected body region. This kind of information is useful background for the development of mechanistic models.

2.5 NCSS Data

The NCSS data consists of accident data obtained from a probability sample of accidents within seven purposively chosen areas. The accidents were sampled from police-reported accidents that involved at least one towed vehicle. During Phase 1, the first fifteen months (January 1977 to March 1978) of NCSS data collection, a police-reported accident was eligible for sampling only if it involved a towed passenger car. In Phase 2 (April 1978 to March 1979) of NCSS data collection a police-reported accident was eligible for sampling if it involved a towed passenger car, light truck or van. Each accident was selected

with a probability which varied with the severity of injury to an occupant of a towed vehicle. Severe accidents were sampled at a higher rate. The sample design within each area assured a random selection of accidents. It guarded against intentional bias of particular types of accidents and provides a probability-based sample of accidents within areas.

The NCSS data structure is hierarchical in nature. For each accident sampled, information was collected for all vehicles in the accident and all occupants in towed vehicles only. Accident level variables describe the environment in which the accident occurred. The vehicle level variables describe the vehicle and the damage done to the vehicle in the accident. Among these variables are the CDC variables and Delta V. Delta V is calculated using CRASH program. Included with Delta V are the longitudinal and lateral components. The occupant level variables contain basic descriptive information about each occupant. Injury information is contained in the Occupant Injury Classifications (OIC) which have associated contact points. In Phase 1 up to three OIC's are coded for an occupant and in Phase 2 this was increased to six. Injury severity information is summarized in the Overall Abbreviated Injury Scale (OAIS).

The missing data rates for the measure of crash severity and injury severity were quite high. The missing data rate for OAIS is approximately 30%. Other measures of injury severity had higher missing data rates. OAIS has a high missing data rate because official medical records were needed to code injury severity. A new variable to represent injury severity was generated using an NCSA algorithm that is documented in Appendix A. The new variable was created using all available information on injury severity collected for NCSS. It classified injuries into two categories: severe (OAIS greater than 2) and non-severe (OAIS less than 3). The missing data rate for Delta V is difficult to calculate. The use of CRASH is restricted to certain types of collisions and missing data rates only apply to these crashes. Information about the suitability of running CRASH for a particular vehicle is not directly available. An estimate of the missing data rate

for Delta V is 30%. The CDC variables have a slightly lower missing data rate.

For the most part, the data used in the Phase 1 and Phase 2 analyses were preliminary versions of the data. These preliminary data sets were missing less than one percent of the final number of NCSS investigated accidents in each phase. The effect of this missing data is thought to be negligible.

2.6 Summary

To briefly review the proceeding material, the mechanistic models seek to describe variation in injury severity on a case-by-case basis in terms of various independent variables. The basis for these models is the physical principles and mechanisms that govern the collision event. The objective is to reflect these relationships. Consequently, the variables and functional form must be consistent with the existing understanding of the physics and biomechanics of the event. In this way, one attempts to choose variables and develop models which reflect true cause and effect relationships.

Population models, on the other hand, seek to define the national accident experience across the levels of various descriptive variables such as collision type and severity. Here, the problem is to estimate the distribution for a larger population from the sample data. No physical models or cause and effect relationships are involved here.

Accident analysis models are a combination of the population and mechanistic models. The objective of the accident analysis models is the estimation of the potential benefit of improved occupant protection systems. These models begin with the population distributions of collision type and severity. The objective is the same here: to define homogeneous subsets of the accident experience for the application of mechanistic models. However, in the accident analysis models, the mechanistic models are modified to reflect the expected effect of improved restraint systems. Ideally, the control variables incorporated in the mechanistic models developed from the accident data would be incorporated in the mechanistic models used in the accident analysis models.

An important aspect of the overall relationship of the mechanistic, population, and accident analysis models is the common and central role played by the selection of the collision types, or modes, and the collision severity variable. The overall success of this effort is measured in terms of the ability of the mechanistic models to predict injury severity. The major difficulty arises from the fact that the levels of the various independent variables were determined by an ongoing social process rather than a deliberate experimental design. Any inferences of cause and effect must be based on theoretical knowledge of the physical principles and mechanisms governing the collision event. The mechanistic models must reflect cause-and-effect relationships if the accident analysis models are to succeed.

3 MECHANISTIC MODELS

The topic of mechanistic models was introduced in Section 2. The objective is to develop statistical models which reflect the physical principles and mechanisms which govern the events of interest. In this case the events of interest are in the area of vehicle crashworthiness. Given that a collision has occurred, the objective is to relate the variables which describe the type and severity of impact, the vehicle, and the occupant to the resulting occupant injuries. It is hoped that a better understanding of the factors which influence the resulting injuries will lead to the development of more crashworthy vehicles. For this expectation to be feasible, the statistical models must reflect the governing physical principles. In current practice, these considerations guide the selection of variables and the formulation of the model.

The basic approach is to subset the data into groups which are felt to be relatively homogeneous with regard to the injury production mechanisms. Model development is then carried out within each subset. In this study, vehicles were initially split into those involved in frontal impacts and those involved in side impacts. During the first half of this study, a preliminary version of the data from the first fifteen months (January 1977 through March 1978) of the NCSS study was used. Subsequently, a final version of the data from the first fifteen months, and a near-final version of the data from the last twelve months (April 1978 through March 1979) were received. The preliminary version of the data from the first fifteen months is referred to as the "Phase 1" data, while the data received for the last twelve months is referred to as the "Phase 2" data. Contact codes were not available in the Phase 1 data. However, the final version of the data from the first fifteen months and the "near-final" version of the data from the last twelve months contained contact codes.

The first subsection presented discusses the analytical techniques used and the logit model in particular. The development of mechanistic models for side impacts is presented in Sections 3.2 and 3.3, the first containing the results obtained with the Phase 1 data and the second containing the Phase 2 and the final results. Modelling of frontal

impacts is covered in Sections 3.4 and 3.5. Again, the Phase 1 results are presented in Section 3.4 and Phase 2 in 3.5. Finally, an overall summary and discussion is provided in Section 3.6.

3.1 Analytical Technique - Logit Analysis

This subsection serves to document the analytical techniques used in the development of models to describe injury severity as a function of crash severity. A description and the use of the logit model is discussed first. Methods for model evaluation are then considered. Finally the effects of sample design and measurement error on the logit model are discussed.

3.1.1 Model Description. This subsection describes the procedure used to analyze specific subsets of the NCSS data set. In the analyses the key dependent variable is a categorical variable with two levels. The specific model that has been used is the logit model. This analysis has been used in various areas of application and is discussed by many authors. A brief discussion of logit analysis follows. For a more detailed discussion see Cox[7], Finney[8], Hanushek and Jackson[9], and Haberman[10]. The statistical package used to compute the logit analysis was the PROBIT function in MIDAS[11]. The PROBIT function uses the method discussed by Aitchison and Silvey[12] modified to include both the probit and logit models.

---

[7] Cox, D.R. (1970), Analysis of Binary Data, Methuen, London.

[8] Finney, D.J.(1971), Probit Analysis, Third Edition, University Press, Cambridge.

[9] Hanushek, Eric A. and Jackson, John E. (1977), Statistical Methods for Social Scientists, Academic Press, New York.

[10] Haberman, Shelley J. (1978), Analysis of Qualitative Data, Volume 1, Academic Press, New York.

[11] Michigan Interactive Data Analysis System (MIDAS), written by Dan Fox and Ken Guire at the Statistical Research Laboratory at the University of Michigan.

[12] Aitchison, J. and Silvey, S.D. (1957), "The generalization of probit analysis to the case of multiple responses", Biometrika 44, pp 131-148.

The logit model postulates a relationship between the probability of observing a category and independent variables. This relationship allows the prediction of the dependent variable given the levels of the independent variables. Since, in this model, the dependent variable is categorical its coded value has no meaning other than to describe which category it is. The notion of prediction must be carefully defined.

It is more natural to think about predicting the probability of observing one of the categories rather than the predicted value for the categorical variable. The data that is collected consists of observations which count the number of times the category appears. This data could be modelled by using the binomial distribution if the probability of a category occurring was the same for all observations. With some categorical variables there is reason to suspect that the probability of a category (for example, injury with severity greater than or equal to 3) is different for each observation. If it is reasonable to believe that these probabilities vary systematically with other variables that can be observed, a mathematical model can be developed to describe this situation.

For example, in the models developed in this project the dependent variable is a dichotomous variable. The two categories described by this variable are non-severe and severe injury. This variable was recoded from the NCSS data using an NCSA developed algorithm described in Appendix A. The assumption is made that the probability of a non-severe injury depends on crash severity and other variables such as crash mode and occupant age. If there is a model to predict the probability of a non-severe injury, the probability of a severe injury is known. The probability of a severe injury is one minus the probability of a non-severe injury. The logit model was used because it provided a general model for predicting the probability of non-severe injury. This model constrains all predicted probabilities to lie between 0 and 1 and the predicted probabilities when graphed as a function of Delta V produces an s-shaped curve increasing with increasing Delta V.

More specifically, the logit model assumes that associated with the dichotomous variable, D, that takes on values 0 and 1, is a continuous

random variable Y. The distribution of Y is the logistic distribution, F^*. F^* is assumed to have a mean of $\underline{\theta}'\underline{X}$ and a variance of 1 where $\underline{\theta}$ is a kx1 vector of parameters and $\underline{X}$ is a kx1 vector of independent variables. Under these assumptions the probability that D is 0 can be expressed as

$$P(D=0|X) = F^*(\emptyset)$$

$$= F[\emptyset - \underline{\theta}'\underline{X}]$$

$$P(D=0|X) = [1 + EXP\ -(\emptyset - \underline{\theta}'\underline{X})]^{-1} \quad (3\text{-}1)$$

where

$\emptyset$ is the threshold,

$\underline{X}$ represents the independent variables,

$\underline{\theta}$ is the vector of unknown coefficients of $\underline{X}$,

F is the logistic distribution with a mean of 0, a variance of 1, and

EXP is the exponential function.

From Equation 3-1 it can be seen that the probability of Category 0 is a distribution function, so the probability is just the cumulative probability up to the threshold, adjusted for the mean, $\underline{\theta}'\underline{X}$, of the underlying distribution of Y. This is shown graphically in Figure 3.1.

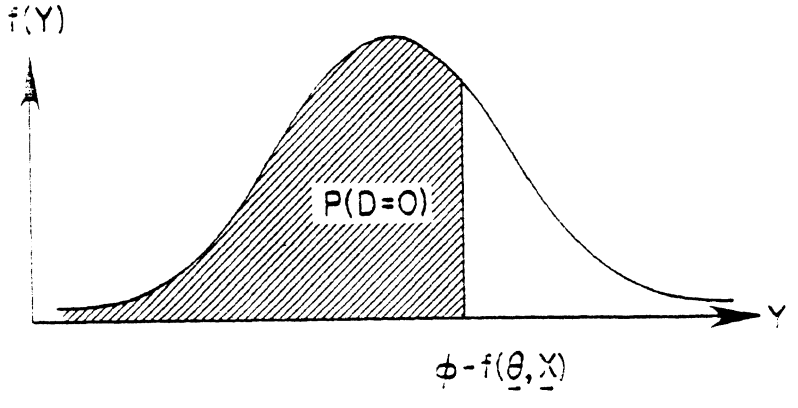

FIGURE 3.1 The Logit Model Probabilities

The basic assumption is that each observation has a binomial distribution with an unknown parameter, P_i, that is specific to each

observation made. The natural logarithm of the likelihood, $L(\emptyset,\underline{\theta})$, is given by

$$(3\text{-}2) \qquad \ln L(\emptyset,\underline{\theta}) = \sum_{i=1}^{n} D_i[\ln P_i] + (1 - D_i)[\ln (1 - P_i)]$$

where

D_i is the value of the dichotomous random variable for the i^{th} observation,

P_i is the $P(D_i = 0)$ defined by Equation 3-1, and

n is the number of observations.

If all the observations were thought to have the same probability that Category 0 would occur, the likelihood given in Equation 3-2 would simplify to the ordinary binomial likelihood with one parameter, P.

This model formulation is slightly more general than the equivalent empirical logit transformation[13] described by

$$\ln \hat{p}_i/(1-\hat{p}_i)$$

where $\hat{p}_i$ is the estimated probability at X_i.

This transformation requires that for every value of $\underline{X}_i$ the proportion in Category 0 be between 0 and 1. This means that there must be at least 2 observations at each specific $\underline{X}_i$ value. This requirement is unnecessary with the model specification described above.

Once the model is thus specified, maximum likelihood theory can be used to estimate the parameters in the model and the variances of the estimates of the unknown parameters. This method of estimation involves an iterative procedure that finds estimates which maximize the binomial likelihood in Equation 3-2 with probabilities as specified in Equation 3-1. It should be noted here that the output from the analysis yields an estimated "regression" equation which when properly transformed yields a prediction of the probability of Category 0 occurring. This predicted probability is given by

$$(3\text{-}3) \qquad \hat{p}_i = [1 + EXP\ -(\hat{\emptyset} - \hat{\underline{\theta}}'\underline{X})]^{-1}$$

---

[13] Cox, D.R. (1970), Analysis of Binary Data, Methuen, London, Chapter 3.

where

$\hat{p}_i$ is the estimated probability of Category 0 for the i^{th} observation,

$\hat{\emptyset}$, and $\hat{\underline{\theta}}$, are the maximum likelihood estimates of $\emptyset$ and $\underline{\theta}$.

3.1.2 Model Development Various methods of exploratory data analysis were used both in choosing subsets within which modelling efforts were thought to be reasonable and in choosing independent variables to use as predictor variables in the regression equation. Those results will be discussed in more detail in the first and second subsections of Sections 3.2 and 3.4. Various statistical tests and measures were used to aid in the choice of the best fitting model. Goodness of fit measures, to be discussed in the next subsection, along with a statistical test of significance, were primarily used to assess the "significance" of a particular variable in the prediction equation.

The test of significance used was a Likelihood Ratio Statistic[14] (LRS), that is generally used to test a null hypothesis that specifies fewer parameters then the alternative. For example consider the following hypothesis:

$$H_0: \; P_i = F(\emptyset)$$

$$H_1: \; P_i = F(\emptyset - \underline{\theta}_1 X_1 - \underline{\theta}_2 X_2)$$

The null hypothesis states that the probability of Category 0 is independent of X_1 and X_2. In general the LRS is defined by

$$\text{LRS} = -2[\ln L(\hat{\emptyset},\hat{\underline{\theta}}|H_0) - \ln L(\hat{\emptyset},\hat{\underline{\theta}}|H_1)] \tag{3-4}$$

and this statistic has an asymptotic chi-square distribution with k degrees of freedom, where k is the number of parameters estimated under the alternative hypothesis minus the number of parameters estimated under the null hypothesis. Thus, to test the hypothesis stated above, the following Likelihood Ratio Statistic is used:

$$\text{LRS} = -2\,[\ln L(\hat{\emptyset}) - \ln L(\hat{\emptyset},\hat{\theta}_1,\hat{\theta}_2)]$$

which has a chi-square distribution with 2 df.

---

[14] Rao, C. R. (1965), Linear Statistical Inference and Its Applications, John Wiley Sons, New York, pp 350-351.

More generally this test can be used to test the additional effect of one variable or group of variables on a model that may have several variables already in it. This test becomes conditional on the variables fixed in the model. That is, new variables that appear not to be significant with one set of variables may appear to be significant with a different set. This situation is similar to regression analysis.

Finally this test statistic, LRS, was used to decide whether the Phase 1 data could be pooled with the Phase 2 data. In a similar context, subsets were evaluated to see if combining data by collapsing some of the subsets was reasonable. Combining data is only reasonable when the model fit on the combined data does not significantly differ from the models fit on the individual subsets. To illustrate the use of the LRS to aid in this decision the problem of combining the Phase 1 and Phase 2 data will be looked at in more detail.

The hypotheses involved in this problem are as follows:

$$H_0: \; L_0(\emptyset,\theta) = L_1(\emptyset,\theta)\cdot L_2(\emptyset,\theta)$$

$$H_1: \; L_1(\underline{\emptyset},\underline{\theta}) = L_1(\emptyset_1,\theta_1)\cdot L_2(\emptyset_2,\theta_2)$$

where

$L_1(\emptyset,\theta)$ is the likelihood of the Phase 1 data under H_0,
$L_2(\emptyset,\theta)$ is the likelihood of the Phase 2 data under H_0,
$L_1(\emptyset_1,\theta_1)$ is the likelihood of the Phase 1 data under H_1,
$L_2(\emptyset_2,\theta_2)$ is the likelihood of the Phase 2 data under H_1,
$(\emptyset,\theta)$ are the parameters associated with H_0, and
$(\emptyset_1,\theta_1,\emptyset_2,\theta_2)$ are the parameters associated with H_1.

The estimates of $(\emptyset,\theta)$ under the null hypothesis are obtained by fitting the logit model on Phase 1 and Phase 2 data. There are only two parameters estimated. These two parameters specify the regression equation in the logit model for the combined data. The logit model is then computed separately on the Phase 1 data and the Phase 2 data to get the four estimates $(\hat{\emptyset}_1,\hat{\theta}_1,\hat{\emptyset}_2,\hat{\theta}_2)$.

The likelihood ratio statistic that tests the null hypothesis that one model is sufficient to describe both the Phase 1 and Phase 2 data sets is given by:

$$LRS = -2 [\ln L_0(\hat{\emptyset},\hat{\theta}) - \ln L_1(\hat{\underline{\emptyset}},\hat{\underline{\theta}})]$$

which, for the example above with only one independent variable, has an asymptotic chi-square distribution with 2 df. For models with k independent variables, the test would have k+1 degrees of freedom.

3.1.3 Goodness of Fit. The Likelihood Ratio Statistic described in the preceding subsection does not tell how much the "regression model" adds to the predictive aspect of the problem. In order to evaluate how well the model predicts, it seems reasonable to calculate for each observation, using the estimated logit model, the probability of the event occurring. The estimate of the predicted probability (predicting Category 0) was given in Equation 3-3. If this probability is greater than one half then the model predicts Category 0 to occur, and if the probability is less than one-half, the prediction would be that Category 0 did not occur. Since the observed data tells which category occurred for each observation, comparing the observed data with the prediction from the model will give some idea about how well the model is predicting. This can be quantified by using as a measure of goodness of fit the percentage of correct predictions.

If the simple binomial model is assumed with a common parameter for each observation, only one probability is estimated. This probability, if it were less than one-half, would lead one to the prediction, for all of the observations, that Category 0 would not occur. If the probability was greater than one-half, the prediction for all observations would be that Category 0 would occur. One simple measure of the predictive power of the model would be to look at the percent correctly predicted using the prediction equation as compared to the percent correctly predicted assuming the regression equation was a constant for all observations.

It was found that the overall percentage of correct predictions was not sensitive enough to detect small changes in the predictive capabilities. In the application of this method in the analysis of NCSS data the percentage correct in the severe injury category was more important than the overall percentage of correct prediction. In evaluating different independent variables for inclusion in the model a significant increase in the percentage of correct predictions for the

severe category was used as a criterion for inclusion of the independent variable in the model.

To define these measures of goodness of fit consider the hypothetical contingency tables described by Figures 3.2 and 3.3.

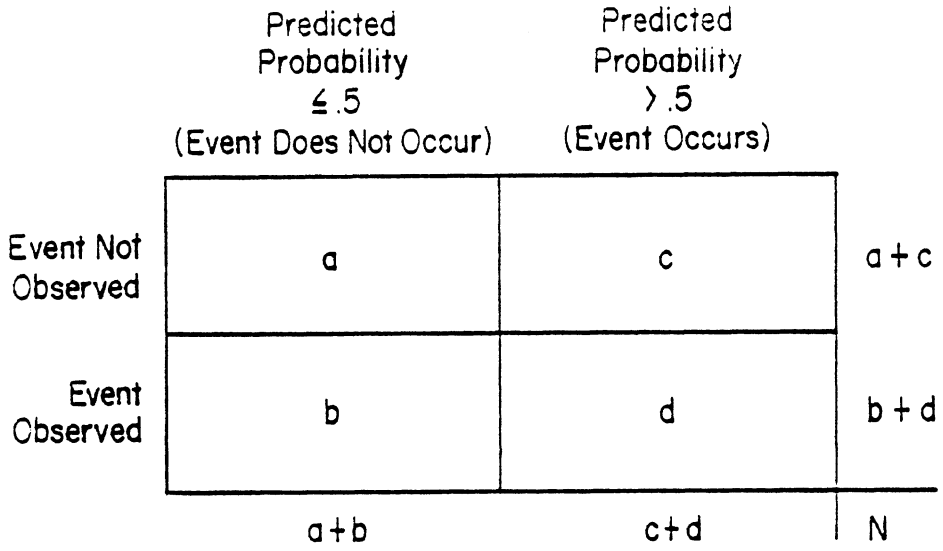

FIGURE 3.2 Contingency Table For Overall Goodness of Fit

The overall percentage of correct prediction is given by

(3-5) Percent Correct Prediction (Overall) = $(a + d)/N$

where a, d, and N are defined in Figure 3.2. The percentage of correct predictions within categories is given by

(3-6) Percent Correct Prediction (Category 0) = $(a_1 + d_1)/N_1$ and

Percent Correct Prediction (Category 1) = $(a_2 + d_2)/N_2$

where a_1, a_2, d_1, d_2, N_1 and N_2 are defined in Figure 3.3.

These measures are all based on the predicted probabilities. These predicted probabilities are dichotomized based on the magnitude of the predicted probability. These predicted probabilities are estimates of the true probability for the observation and hence are variable

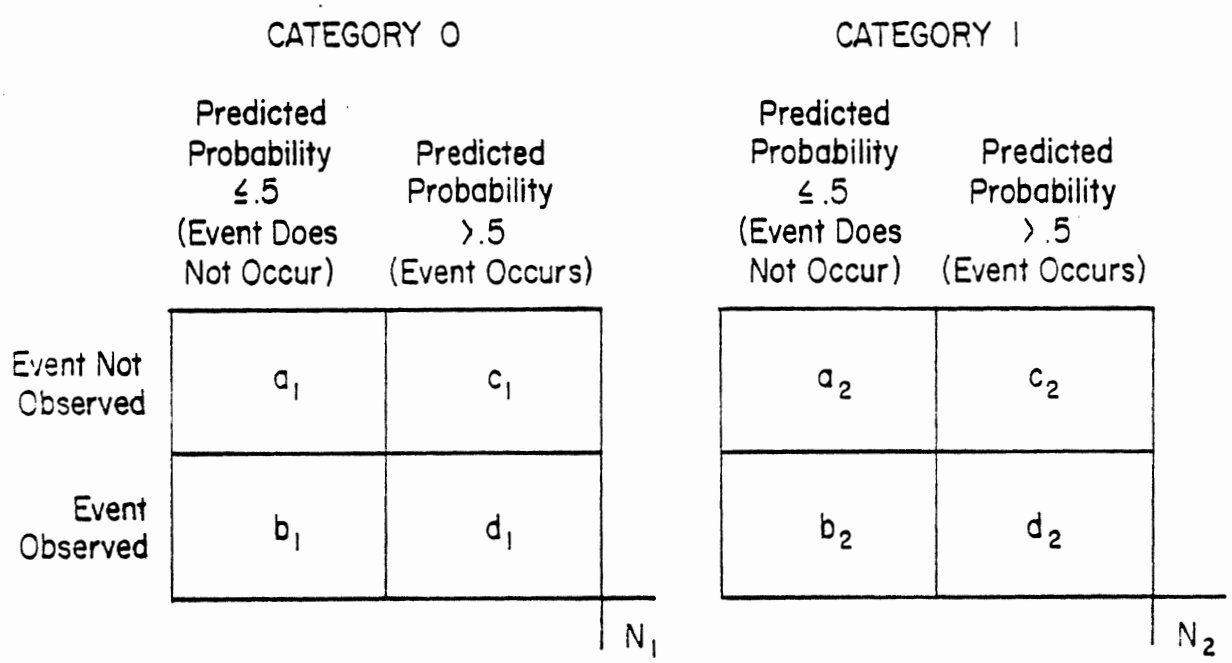

FIGURE 3.3 Contingency Table For Goodness of Fit By Categories

themselves. Their variability may be large enough to make uncertain which category a predicted probability of .4 or .6 should fall into. Histograms of the predicted probabilities from a model for each category were examined to see what distribution these estimated probabilities had. This information was used to evaluate the model and used in the analysis of mispredictions (outliers). For reference, Figure 3.4 gives a hypothetical distribution for the predicted values within categories. In the figure, Category 0 is non-severe injuries and Category 1 is severe injuries. The left hand distribution shows the model predictions for all cases of non-severe injuries. For instance one case has a predicted probability of .1 while eight cases have a predicted probability of 1.0. For those observations in Category 1 it is expected that the predicted probability (predicting Category 0) would tend to be small, no greater than .5. For those observations in Category 0, the predicted probabilities (predicting Category 0) would all tend to be close to 1. This is the case shown in the figure. Note that the model would mispredict some cases in each category, the eight non-severe cases with a predicted probability of less than .5 and the nine severe cases with the predicted probability greater than .5.

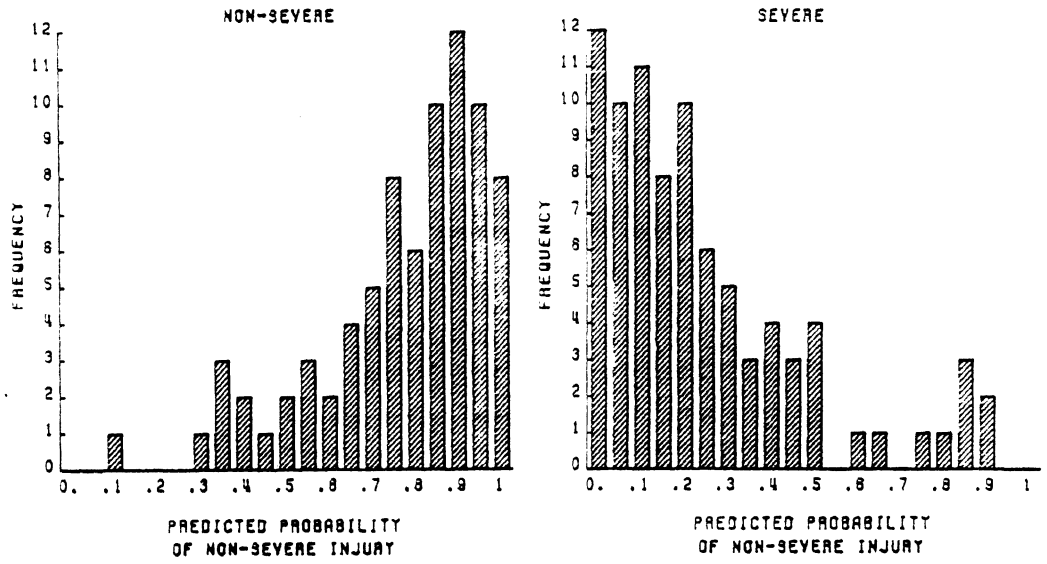

FIGURE 3.4 Hypothetical Histograms of Estimated Predicted Probabilities

Those estimated predicted probabilities that deviate from this expected pattern could be considered to be extreme values. These extreme values (or mispredictions) can be examined to see if there is any common element. Upon identification of such a factor a new variable could be added to the model to improve the predictability of the model.

3.1.4 Confidence Intervals. The model assumes that

$$(3\text{-}7) \qquad P(\underline{X}) = [1 + EXP\ -(\emptyset - \underline{\theta}'\underline{X})]^{-1}$$

In this formulation $P(\underline{X})$ is the probability associated with the value of $\underline{X}$, where $\underline{X}$ is a k x 1 vector of observations on k independent variables, $\underline{\theta}'$ is a 1 x k vector of parameters (coefficients of $\underline{X}$), $\emptyset$ is the threshold parameter and EXP represents the natural exponential function. Maximum likelihood estimates and the estimated variances and covariances of these estimates are obtained from the computer program. These estimated variances and covariances are used to calculate the variance of the predicted probability, $\hat{p}(\underline{X})$ associated with a value of $\underline{X}$. The

estimate of $P(\underline{X})$ is given by substituting in Equation 3-7 maximum likelihood estimates for $\emptyset$ and $\underline{\theta}$.

A Taylor series approximation[15] method can be used to obtain approximations to the expected value and variance of the estimates of the probabilities generated by the logistic model. Using this technique it can be shown that

$$(3\text{-}8) \qquad E\,\hat{p}(\underline{X}) = P(\underline{X})$$

The estimates, $\hat{\emptyset}$ and $\hat{\underline{\theta}}$, are maximum likelihood estimates and are therefore consistent. Thus, $\hat{p}(\underline{X})$ is also a consistent estimate of $P(\underline{X})$. The theoretical variance of a predicted probability at $\underline{X}$ is given by

$$(3\text{-}9) \qquad \text{Var}\,P(\underline{X}) \doteq P(\underline{X})^2\,[1 - P(\underline{X})]^2\,\text{Var}\,[\emptyset - \underline{\theta}'\underline{X}]$$

where $P(\underline{X})$ is defined as in Equation 3-7. The variance on the right hand side of Equation 3-8 can be rewritten as

$$(3\text{-}10) \qquad \text{Var}\,[\emptyset - \underline{\theta}'\underline{X}] = \text{Var}\,\emptyset + \underline{X}'\text{Var}(\underline{\theta})\underline{X} - 2\text{Cov}\,(\emptyset,\underline{\theta}')\underline{X}$$

where Var $\emptyset$ and Var $\underline{\theta}$ are the theoretical variance and covariance matrices of $\emptyset$ and $\underline{\theta}$ respectively, and Cov $(\emptyset,\underline{\theta}')$ is a vector of true covariances between the theoretical threshold and the coefficients of the independent variables.

To estimate this fairly complicated expression for the variance, the maximum likelihood estimates for the parameters that were obtained in the model fitting procedure will be used. The resulting variance estimate is given by

$$(3\text{-}11) \qquad \hat{\text{Var}}\,\hat{p}(\underline{X}) = \hat{p}(\underline{X})^2[1 - \hat{p}(\underline{X})]^2[\hat{\text{Var}}\hat{\emptyset} + \underline{X}'(\hat{\text{Var}}\hat{\theta})\underline{X} - 2\hat{\text{Cov}}(\hat{\theta},\hat{\emptyset}')\underline{X}].$$

It is appropriate to note here that this variance formula represents the variance of a prediction at a given point $\underline{X}$ and with it

---

[15]Kendall, M. G. and A. Stuart, The Advanced Theory of Statistics (1958), Volume 1, Chas Griffin Co. Ltd. pp 232-233.

an approximate confidence interval can be formed around the prediction at that $\underline{X}$. Calculating these confidence intervals at many $\underline{X}$ values is not equivalent to using techniques that produce simultaneous confidence intervals. The resulting "confidence band" using this variance formula will not be equivalent to a "simultaneous set of confidence intervals."

These derivations assume the most general situation for the logit model in terms of the number and type of independent variables there are in the model. At this point three special cases will be considered:

1) One independent variable that is continuous,

2) Two independent variables that are continuous, and

3) Dummy variables with continuous independent variables.

With only one continuous variable in the regression equation the estimated variance of a predicted probability at X is given by

$$(3\text{-}12) \qquad \hat{\text{Var}}\ \hat{p}(X) = \hat{p}(X)^2[1 - \hat{p}(X)]^2[\hat{\text{Var}}\ \hat{\emptyset} + X^2\ \hat{\text{Var}}\ \hat{\theta} - 2X\ \hat{\text{Cov}}\ (\hat{\emptyset},\hat{\theta})].$$

Using this estimate of the variance the approximate 95% confidence interval[16] for the true probability is given by

---

[16]This confidence interval is based on an estimated variance that was obtained with an approximation. The method used to develop this approximation involved expanding the functional form of the probability (the logistic function) in a Taylor Series expansion. Only the linear term of the expansion was used to approximate the probability. This linearized version of the logistic function was used to derive an approximation to the variance of the probability. There are conditions when this is not a good approximation. One specific case is when the probability is extreme, close to 0 or 1. In this case the approximation yields a large estimated variance and the confidence intervals may exceed 0 and 1. For examples see Figures 3.18, 3.19, 3.34, 3.35, 3.51, 3.52, and 3.65. It should be noted that probabilities close to 0 or 1 imply almost all of the data is in one category or the other. This observation alone might indicate that the modelling of this data may be unstable. An alternative method would be to calculate a confidence interval for the logarithm of the odds ratio which is exactly linear in the independent variables. To get a confidence interval for the probability a transformation needs to be done. This method constrains any confidence interval between 0 and 1 and can be asymmetric. When the probability is not extreme these methods provide approximately the same confidence intervals.

(3-13) $\hat{p}(X) \pm \sqrt{1.96 \widehat{\text{Var}}\, \hat{p}(X)}.$

Since there is only one independent variable, $\hat{p}(\underline{X})$ and its approximate confidence interval can easily be graphically displayed.

When there are two independent continuous variables in the regression model plotting $\hat{p}(X_1,X_2)$ and plotting the approximate confidence interval around the predicted probability at (X_1,X_2) becomes difficult because of the three dimensional nature of the problem. One method to graphically represent $\hat{p}(X_1,X_2)$ and its variance is to use contour plots to graphically represent $\hat{p}(X_1,X_2)$ and the $\widehat{\text{Var}}\, P(X_1,X_2)$. Although this type of representation of $\hat{p}(X_1,X_2)$ and its variance may be informative, it would be difficult to associate $\hat{p}(X_1,X_2)$, for a particular value of X_1 and X_2, and its estimated variance.

An alternative method reduces the problem to the two dimensional situation. Here the graphical representation used in the one continuous variable case is also used but now a series of graphs need to be presented for each model. If one of the variables, X_2, is fixed at a particular value then $\hat{p}(X_1,X_2)$ can be graphed as a function of X alone and its variance calculated by

$$\begin{aligned}(3\text{-}14)\quad \widehat{\text{Var}}\, \hat{p}(X_1X_2) = \hat{p}(X_1X_2)^2\,[1 - \hat{p}(X_1X_2)]^2[&\widehat{\text{Var}}\, \hat{\emptyset} + X_1^2\, \widehat{\text{Var}}\, \hat{\theta}_1 \\ &+ X_2^2\, \widehat{\text{Var}}\, \hat{\theta}_2 - 2X_1\, \widehat{\text{Cov}}\,(\hat{\emptyset},\hat{\theta}_1) - 2X_2\, \widehat{\text{Cov}}(\hat{\emptyset},\hat{\theta}_2) \\ &+ 2X_1X_2\, \widehat{\text{Cov}}\,(\hat{\theta}_1,\hat{\theta}_2)]\end{aligned}$$

where

$\hat{\theta}_1$ is the coefficient of X_1 in the model and
$\hat{\theta}_2$ is the coefficient of X_2 in the model.

Now for each value of X_2, an approximate confidence interval can be formed similar to the one described in Equation 3-13 and indicated on the graph of $\hat{p}(X_1,X_2)$ as a function of X_1.

The only problem that remains is the choice of X_2, that is, the number of different values of X_2 that are to be considered. This can not be decided independently of the particular model under study. Choices of X_2 will depend on how sensitive the model is to X_2. In some situations, although probably few, evaluation of $\hat{p}(X_1,X_2)$ only at some

"average" value of X_2 will be required. More likely than not, besides an "average" value at X_2, a value of both extremes will be deemed necessary.

Dummy variables are usually incorporated into a model to represent a categorical variable or several categorical variables. Usually there are (I-1) dummy variables, each variable associated with one of the levels of a categorical variable with I levels that are added to a model. These dummy variables are coded 1 if the category occurs for that individual and are zero otherwise. A typical model involving dummy variables, W_1, W_2, and W_3, would be

$$P(X,W_1,W_2,W_3) = F(\emptyset - \theta_1 W_1 - \theta_2 W_2 - \theta_3 W_3 - \theta_4 X)$$

where X is continuous, W_1, W_2, and W_3 represent a categorical variable (with 4 levels), and F is the logistic distribution. These dummy variables, in effect, subtract the constant θ_1 if Category 1 occurs, the constant θ_2 if Category 2 occurs, and θ_3 if category 3 occurs. This model can be rewritten as three different models given that a particular category occurs:

$$\hat{p}(X_1,W_1,W_2,W_3) = F(\hat{\emptyset}-\hat{\theta}_1-\hat{\theta}_4 X) \text{ if Category 1 occurs,}$$

$$\hat{p}(X_1,W_1,W_2,W_3) = F(\hat{\emptyset}-\hat{\theta}_2-\hat{\theta}_4 X) \text{ if Category 2 occurs,}$$

and $$\hat{p}(X_1,W_1,W_2,W_3) = F(\hat{\emptyset}-\hat{\theta}_3-\hat{\theta}_4 X) \text{ if Category 3 occurs.}$$

These estimated probabilities, for each category, can be graphically represented with the approximate confidence intervals. The variance estimate is given by Equation 3-11 where $\hat{\emptyset}$ is replaced by the estimate $(\hat{\emptyset}-\hat{\theta}_i)$ for each Category i.

Models involving dummy variables and one continuous variable will involve two graphical representations for each dummy variable in the model. If there are two continuous variables, for each of these graphs three or possibly more graphical representations will be required.

With models involving three or more continuous variables, graphical representation becomes infeasible because of the number of different possibilities one might have to consider. The general variance formula is given by Equation 3-11 and for any specific point the estimated

probability and its estimated variance can be calculated and an approximate 95% confidence interval formed.

3.1.5 Measurement Errors. The model specification assumes that the mean of the underlying continuous variable, Y, in the logit model is $\underline{\theta}'\underline{X}$. The vector $\underline{X}$ is assumed to be a vector of constants, that is, $\underline{X}$ is not itself random. For most of the variables in NCSS this assumption may be reasonable. There is one notable exception: Delta V and its longitudinal and lateral components. This subsection investigates analytically the effect of the coefficients in the logit model when the independent variable is subject to error. This analysis is related to work in progress by Smith[17].

In the context of the model described in Section 3.1.1 suppose that

$$Y = \underline{\theta}'\underline{X} + \theta_0 X_0 + e \tag{3-15}$$

where

e has a distribution function F^* with variance of 1,

$\underline{X}$ and X_0 are independent variables observed without error, and

$\underline{\theta}, \theta_0$ are unknown parameters.

Then as before

D = 0 if Y is less than Ø

D = 1 if Y is greater than Ø

so that

$$P(D = 0) = F^*[\emptyset]$$

$$= F[\emptyset - \underline{\theta}'\underline{X} - \theta_0 X_0] \tag{3-16}$$

where F is a distribution function with mean 0 and variance 1. The probabilities in Equation 3-16 were used in the binomial likelihood and maximum likelihood estimates were obtained for $(\underline{\theta}, \theta_0)$ and $E\,\hat{\underline{\theta}} = \underline{\theta}$ and $E\hat{\theta}_0 = \theta_0$.

Now suppose that it is reasonable to assume that the measurement error in X_0 is such that it only adds variability to the true value of X_0. That is, the measurement process does not by nature produce a

---

[17] Smith, David W., Personal Communication, April 1980.

biased measurement for X_0. With measurement error, X^* will be the observed value for X_0 where

(3-17) $$X^* = X_0 + f$$

where

X_0 is the "true value of the independent variable and
f is a random variable with mean 0 and variance s^2.

This measurement error model assumes that f is independent of e, the error in Y. Using this model for the measurement error in X_0 we can rewrite Equation 3-15 as

(3-18) $$Y = \underline{\theta}'\underline{X} + \theta_0(X^* - f) + e$$

$$= \underline{\theta}'\underline{X} + \theta_0 X^* + e^*$$

where $e^* = e - \theta_0 f$ is a random variable with mean 0 and variance $t^2 = (\theta_0^2 s^2 + 1)$. In estimating the probability of Category 0 with the observed X^*, rather than the true X_0, as a variable, we have

(3-19) $$P(D = 0|X^*) = F[(\emptyset - \underline{\theta}'\underline{X} - \theta_0 X^*)/t]$$

where t is $(\theta_0^2 s^2 + 1)^{\frac{1}{2}}$.

When the probabilities given by Equation 3-19 are used in the binomial likelihood to define the likelihood, the expected value of the estimates that are obtained are given by

$$E\,\hat{\emptyset} = \emptyset(\theta_0^2 s^2 + 1)^{-\frac{1}{2}},$$

$$E\,\hat{\underline{\theta}} = \underline{\theta}(\theta_0^2 s^2 + 1)^{-\frac{1}{2}}, \text{ and}$$

$$E\,\hat{\theta}_0 = \theta_0(\theta_0^2 s^2 + 1)^{-\frac{1}{2}}.$$

From these equations it can be seen that the maximum likelihood estimates for a logit model where there is measurement error (a random error model) will be biased. The estimates will be unbiased only if the variance specified by the model for measurement error is 0, that is when

no measurement error is present. Since θ_0^2 and s^2 will always be positive the effect of this type of measurement error is to underestimate the magnitude of the true value of all the coefficients in the logit model by a factor related to the magnitude of the measurement error.

3.1.6 Sampling Problems. In the development of mechanistic models ideally it would be best to do a controlled experiment. Using this type of design the attempt would be made to get observations over the complete range of independent variables. Sample sizes could be controlled to obtain nearly equal sample sizes within all of the cells defined by the independent variables of interest.

The data obtained by NCSS were not equally balanced over key independent variables. Even though severe accidents were oversampled, many more accidents with low crash severity are included than accidents with high crash severity. The distribution of occupants by age was representative of the driving population and older occupants appeared less frequently than younger drivers.

It was not known how sensitive the model would be to data that was not fairly evenly distributed across the cells defined by the independent variables. This was empirically investigated during the analysis of the side impacts in the Phase 1 data. A model was chosen with two continuous variables, the measure of crash severity, Delta V and Occupants' Age. The side impacts were then categorized by Delta V and Occupants' Age and equal samples were drawn from each category. The model was then estimated using the "new" sample of data. The model based on all side impacts and the "new" balanced sample model did not differ substantially. The specific analysis is described in Section 3.2.3.

3.2 Preliminary Analytical Results for Phase 1 Data - Side Impacts

This section presents preliminary work to develop mechanistic models for side impacts. The analysis reported in this section was carried out on a preliminary version of the data from the first fifteen months of NCSS (January 1977 through March 1978).

The dependent variable used in this analysis is an NCSA-generated variable called NEWOAIS3. Basically, NEWOAIS3 is two-level categorical variable that has the value 0 if the OAIS[18] is 0-2, and is 1 if the OAIS is 3 or greater. Other injury information in the file is used by the NCSA algorithm, where possible, to generate a value for NEWOAIS3 when OAIS is missing. As a result, NEWOAIS3 has less missing data than OAIS. Throughout this presentation, injuries coded NEWOAIS3=0 (OAIS=0-2) will be referred to as "non-severe," and injuries coded NEWOAIS3=1 (OAIS=3+) as "severe."[19] The goal of this analysis is to predict the probability of a non-severe injury as a function of various relevant variables. The probability of a severe injury is simply one minus the probability of an non-severe injury.

The initial modelling efforts were carried out separately for several subsets of the side impacted vehicles. The selection of these subsets is described in the first subsection. Examination of candidate independent variables is covered in the second subsection. Modelling results and model evaluation are presented in the third and fourth subsections, while the last subsection describes the final models for the Phase 1 data.

3.2.1 Defining Subsets. This analysis included only occupants of case vehicles (towed due to crash damage) involved in two-vehicle collisions. The subsetting of this group is primarily based on the

---

[18] OAIS refers to the overall abbreviated injury score as defined in The Abbreviated Injury Scale, (1976 Revision) American Association for Automotive Medicine, Morton Grove, Illinois.

[19] For the algorithm used in creating NEWOAIS3 see Appendix A.

Collision Deformation Classification[20] variable. The third character of this variable describes the General Area of Damage (GAD). A GAD of L (Left) or R (Right) was used to define a side impact. For GAD=Left or Right, injuries to occupants in two-vehicle collisions represent 71% of the accidents while case vehicles account for 80% of the vehicles in the file. Subsets at various levels of detail based essentially on Specific Horizontal Area (SHL) and seat position, were examined with a view to obtaining analytical cells which were sufficiently homogeneous and yet contain adequate data to yield reasonable modelling results. It was also essential for the subsetting to yield a common range of the key variables under consideration so as to permit a meaningful comparison of the modelling results. The following subsetting was selected:

1. Passenger-compartment damage (SHL=D,P,Y,Z. and occupants on the same side as the impact; to be referred to as Near PCD

2. Passenger-compartment damage and occupants on the opposite side to the impact; to be referred to as Far PCD

3. No passenger-compartment damage (SHL=F+B. and occupants on the same side as the impact; to be referred to as Near NPCD

4. No passenger-compartment damage and occupants on the opposite side to the impact; to be referred to as Far NPCD

5. SHL=F,B,D,P,Y,Z and occupants on the same side as the impact; to be referred to as All Near

6. SHL=F,B,D,P,Y,Z and occupants on the opposite side to the impact; to be referred to as All Far. Subsets 5 and 6 are made up of subsets 1 and 3, and subsets 2 and 4 respectively.

"Occupants" refers to drivers and passengers in both the front and the back seats. Table 3.1 shows the number of occupants for each subset. Also shown is the number of cases with valid Delta V and CDC Extent codes, and the number of severe and non-severe injuries.

3.2.2 Examination of Independent Variables. Some of the potential independent variables which were initially investigated were:

1. Delta V

---

[20]"Collision Deformation Classification--SAE Recommended Practice J224a," SAE Handbook, 1980 Ed. (Warrendale, Pa.: Society of Automotive Engineers, 1980), pp. 34.109-34.113.

TABLE 3.1

Number Of Cases Valid
For Specific Variables

Phase 1 Data - Side Impacts

	Near PCD	Far PCD	Near NPCD	Far NPCD	ALL NEAR	ALL FAR
Total	641	616	354	340	995	956
Non-Severe	370	431	279	290	649	721
Severe	140	55	16	7	156	62
Delta V	423	421	232	223	655	644
CDC Extent	641	616	354	340	995	956

2. Vertical Location of Damage (CDC).

3. Damage Distribution Type (CDC).

4. CDC Extent

5. Vehicle Weight

6. Object Contacted (CDC).

7. Principal Direction of Force (CDC).

Comparison of the six subsets over these independent variables reveals the following:

Delta V. The range of Delta V, the mean and the standard deviation within these six subsets are shown in Table 3.2. When the examination of Delta V was based on each category of SHL, much higher values of average Delta V were found to be associated with SHL=D (Distributed).

Vertical Location of Damage (CDC). For the subsets with passenger compartment damage, 85% of the cases are in "Below Glass" category and some 12% in "All" category. For the subsets with no

TABLE 3.2

Delta V For The Six Subsets

Phase 1 Data - Side Impacts

Group	Sample Size	Range Delta V	Mean	S.D.
Near PCD	423	3-63	15.5	7.9
Far PCD	421	3-63	15.8	8.6
Near NPCD	232	2-42	9.7	5.9
Far NPCD	223	2-36	9.3	5.6
ALL NEAR	655	2-63	13.4	7.7
ALL FAR	644	2-63	13.7	8.2

passenger compartment damage, about 99% of cases are in the "Below Glass."

Damage Distribution Type (CDC). For all subsets, the most frequent damage distribution is "Wide", representing 90% or more of the cases. For the subsets with passenger compartment damage, "Sideswipe" represents 6% to 7% of the cases while for the no passenger compartment damage subsets, "Corner" accounts for 7% of the cases.

CDC Extent. For all subsets, the distribution of CDC extent codes are similar - CDC extent codes 2 and 3 represent 75% to 80% of the cases. However, when the examination of CDC extent was based on each category of SHL, the distribution of CDC extent codes for SHL=D was found to be more widely distributed from one extreme to the other than other categories; only 50% to 60% of the cases were made of CDC extent codes 2 and 3.

Vehicle Weight. For all subsets, the distribution of vehicle weights are similar with the common range being 1300 pounds to 4000 pounds.

Object Contacted (CDC). For all subsets, the most common object contacted are another passenger car (over 80%), truck (11%), tractor-trailer (2%) and train (1%).

Direction of Principal Force. Because of the angular velocity associated with the side collision, a new variable based on CDC direction was created to be used in conjunction with Delta V in explaining the occurrence of injury. This new variable is essentially a measure of θ, the angle between the Principal Direction of Force (CDC Direction) and the lines parallel to the lateral axis of a vehicle. The magnitude of this angle is one of the factors influencing the magnitude of the induced angular velocity. This new variable was created by collapsing the CDC Direction variable into five levels. These levels were characterized by the values of θ (the angle between the direction of impact and the lines parallel to the lateral axis of a vehicle) and the location of impact (whether impact is in the front half or the rear half of a vehicle). The five levels were (1 o'clock and 11 o'clock), (2 o'clock and 10 o'clock), (3 o'clock and 9 o'clock), (4 o'clock and 8 o'clock) and (5 o'clock and 7 o'clock).

3.2.3 Model Estimation. The modelling technique used was the logit regression, the technical aspect of which is discussed in Section 3.1. Univariate models are discussed in the first portion of this section. Next, multivariate models are considered. A final subsection describes the result of a subsampling to achieve a more uniform distribution of cases across the independent variables. Model evaluation and examination of residuals is largely reserved for Section 3.2.4 which follows.

Univariate Models. The initial univariate models explored the relationship of Delta V (total), CDC Extent, and the lateral component of Delta V with injury severity (as indicated by NEWOAIS3). It was anticipated that, for the side collisions, the Direction of Force might be significant in explaining injury severities, especially when used in conjunction with Delta V. Actually, within the six subsets previously defined, Delta V and the Direction of Force, should take into account the effects, if any, of rotation of the vehicle on injury. The subsets, based on Specific Horizontal Location (SHL), signify the location of impact in terms of the distance from the vehicle's lateral axis to the point of impact. The combination of Delta V and Direction of Force represents the component of Delta V that was

parallel to the vehicle's lateral axis (Lateral Delta V). The induced angular velocity of the impacted vehicle is a function of both Lateral Delta V and the location of impact relative to the vehicle's central axis.

The lateral component of Delta V has a magnitude which equals Delta V x cosθ where θ is the angle between the direction of force and the line parallel to the vehicle's lateral axis.

The modelling of injury severity by Lateral Delta V resulted in a better goodness of fit than that of injury severity by Delta V. For the six subsets, the results are shown in Table 3.3. These results clearly indicate that Lateral Delta V is a better predictor of injury than Delta V.

Two univariate models, one having Lateral Delta V as the independent variable and the other having CDC Extent as the independent variable may be compared. CDC Extent predicted injuries as well as Lateral Delta V. Lateral Delta V and CDC Extent are expected to be highly correlated; actually in the cause/effect context, CDC Extent can be thought of as the "effect" of Delta V just as injury is an "effect" of Delta V. The high correlation between Delta V (and therefore Lateral Delta V) and CDC Extent should exclude one of them from the model. Lateral Delta V was considered to be more the "cause" of injury than CDC Extent. Therefore Lateral Delta V was retained as one of the independent variables.

Other univariate models, each with injury-severity as the dependent variable but with a different independent variable, were tested. Such variables are:

Vehicle Variables:	Vehicle Weight
	Total Vehicle Weight
	Object Contacted
	Rural/Urban
Occupant Variables:	Age
	Height
	Occupant Weight
	Sex

The univariate models with each of the above-mentioned as the individual variables did not yield meaningful estimated models of

TABLE 3.3

Comparison of The Univariate Models with Lateral Delta V, Delta V and CDC Extent as the Independent Variable

Phase 1 Data - Side Impacts

Subset	Sample Size		Overall % Correct			Non-Severe % Correct			Severe % Correct		
	Non-Severe	Severe	Lat. Delta V	Delta V	CDC Ext.	Lat. Delta V	Delta V	CDC Ext.	Lat. Delta V	Delta V	CDC Ext.
Near PCD	239	95	78	72	80	95	90	95	36	27	40
Far PCD	287	38	91	90	91	98	97	98	42	37	35
Near NPCD	183	9	96	96	95	100	100	100	11	11	6
Far NPCD	191	3	99	99	98	100	100	100	0	0	0
ALL NEAR	414	104	83	79	83	97	94	93	26	25	40
ALL FAR	478	41	94	93	93	99	98	99	39	27	16

reasonable goodness-of-fit measures. However, they were likely to be more significant in explaining injury severity in the presence of Delta V.

Vertical Location of Damage (CDC) and Damage Distribution Type (CDC) were not tested in the modelling because the data was found to be highly concentrated in only one or sometimes two levels of those variables.

Multivariate Models. The discussion so far has pointed to a model which has Lateral Delta V as a primary independent variable while other vehicle and/or occupant variables can be added into the model to improve its predictive capability.

When the model incorporates more than one independent variable, the significance of the additional independent variable may be determined as follows:

1. A statistical test, based on a Likelihood Ratio Statistic (LRS) which has a chi-square distribution, tests the effect of that variable in the model.

2. The goodness of fit of the new model may be compared to that of the existing model. The goodness of fit gives the percentage of cases correctly predicted by the model for both the low-severity injuries and the high-severity injuries.

The candidates for independent variables are given in Table 3.4. Ideally, for an additional variable to be considered significant in the model, both criteria should be met. However, the first criterion can only be met if the two data sets which were being compared are of the same size. This requirement is not necessary for the second criterion. Also, it is not uncommon for an additional variable to show a significant LRS without improving the model's predictive capability at all. Hence for any variables to be included in the model, improvement in the goodness of fit of the model by such variables is necessary, not just a significant value for the LRS.

TABLE 3.4

Candidates For Independent Variables

1.	Delta V
2.	Lateral Delta V
3.	CDC Extent
4.	Vehicle Weight
5.	Vehicle Weight + Occupant Weight + Cargo Weight
6.	Object Contacted
7.	Rural/Urban
8.	Age
9.	Occupant Height
10.	Occupant Weight
11.	Sex
12.	Intrusion Location
13.	Ejection
14.	Restraint Usage
15.	Injury Type
16.	Body Region
17.	Contact Point

The independent variables , which were found to be significant at this stage were Lateral Delta V and Age. The estimated models for the six subsets are as follows:

Estimated Models with Lateral Delta V and Age

Near PCD (N=325, LRS=77.47, DF=2)

(3-20) $\hat{p}_i = F(2.2974 - 0.1090X_1 - 0.0124X_2)$

Far PCD (N=323, LRS=84.22, DF=2)

(3-21) $\hat{p}_i = F(3.4794 - 0.1199X_1 - 0.0100X_2)$

Near NPCD (N=190, LRS=14.4, DF=2)

(3-22) $\hat{p}_i = F(3.5238 - 0.0724X_1 - 0.0278X_2)$

Far NPCD (N=193, LRS=4.73, DF=2)

(3-23) $\hat{p}_i = F(3.9014 - 0.1078X_1 - 0.0123X_2)$

All Near (N=515, LRS=117.07, DF=2)

(3-24) $\hat{p}_i = F(2.5813 - 0.1150X_1 - 0.0132X_2)$

All Far (N=516, LRS=107.13, DF=2)

(3-25) $\hat{p}_i = F(3.6480 - 0.1246X_1 - 0.0100X_2)$

where

$\hat{p}_i$ is the estimated probability of a non-severe injury,
F is the logistic function,
X_1 is Lateral Delta V,
X_2 is Age, and
LRS is the Likelihood Ratio Statistic.

The goodness of fit for these bivariate models is shown for each subset in Table 3.5. While the overall percent correct prediction is fairly good, the percent correct prediction for the severe injuries is, at best, not quite 50% for the Far PCD group and is much worse for the NPCD (non-passenger compartment damage) groups. The histograms of $\hat{p}_i$ for the six subsets are shown in Figures 3.5 through 3.10. Each figure, representing a model for each subset, consists of a pair of histograms, one for the non-severe cases (or cases which had the OAIS codes of 0-2) and the other for the severe cases (cases which had the OAIS codes of 3-6). Both histograms have the same axes, one representing the

estimated probabilities of non-severe injuries ($\hat{p}_i$) at the 0.05 interval and the other the number of cases with the particular $\hat{p}_i$ values. In all subsets, while the histograms look quite reasonable for the non-severe injuries, the histograms for the severe injuries clearly reflect the relatively poor percent correct prediction of the severe cases.

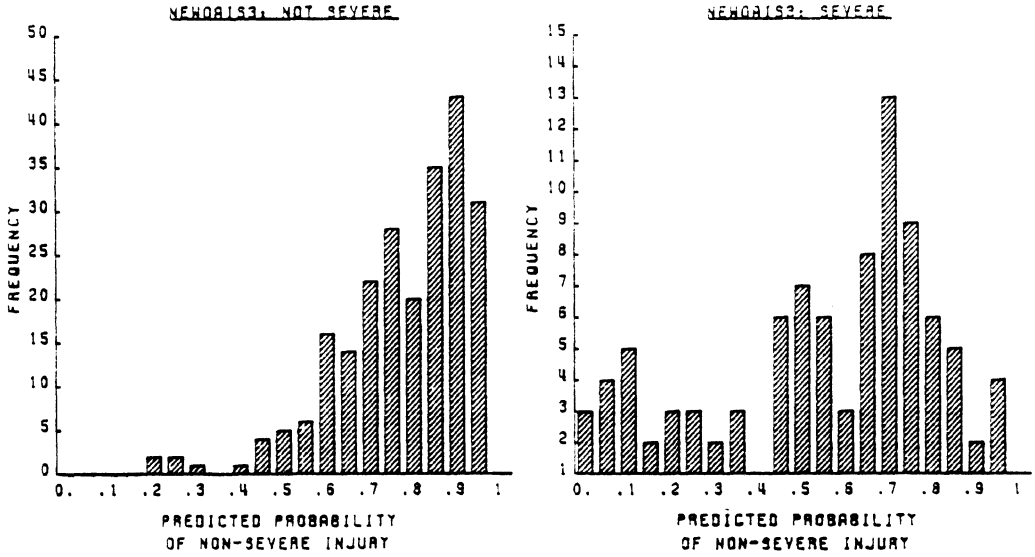

FIGURE 3.5 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For Near PCD Phase 1 Data - Side Impacts

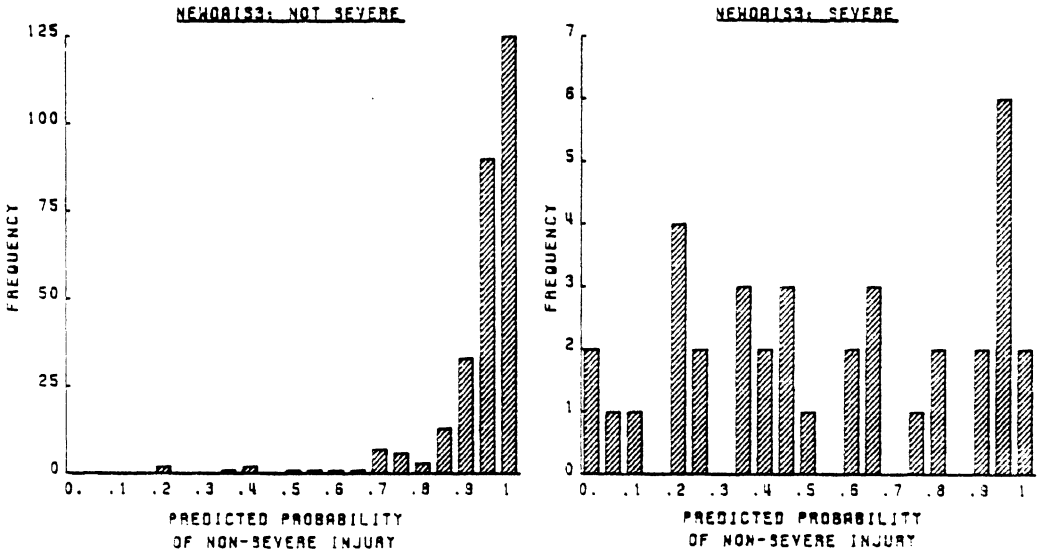

FIGURE 3.6 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For Far PCD Phase 1 Data - Side Impacts

Figure 3.11 shows the logistic curves for the four subsets as given by Equations 3-20 through 3-23. These curves show how the estimated probability of a severe injury varies with Delta V values while Age is

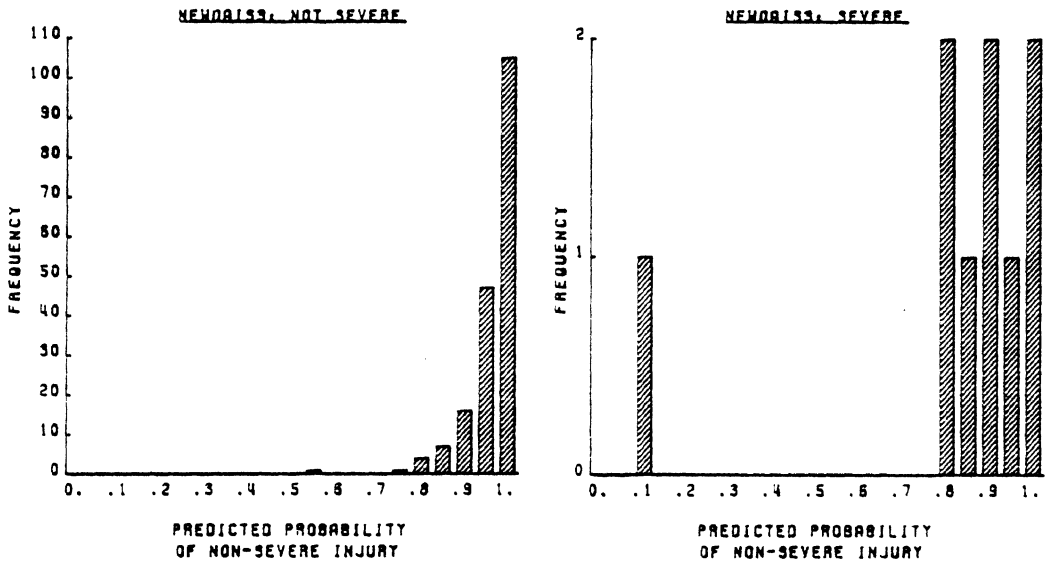

FIGURE 3.7 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For Near NPCD Phase 1 Data - Side Impacts

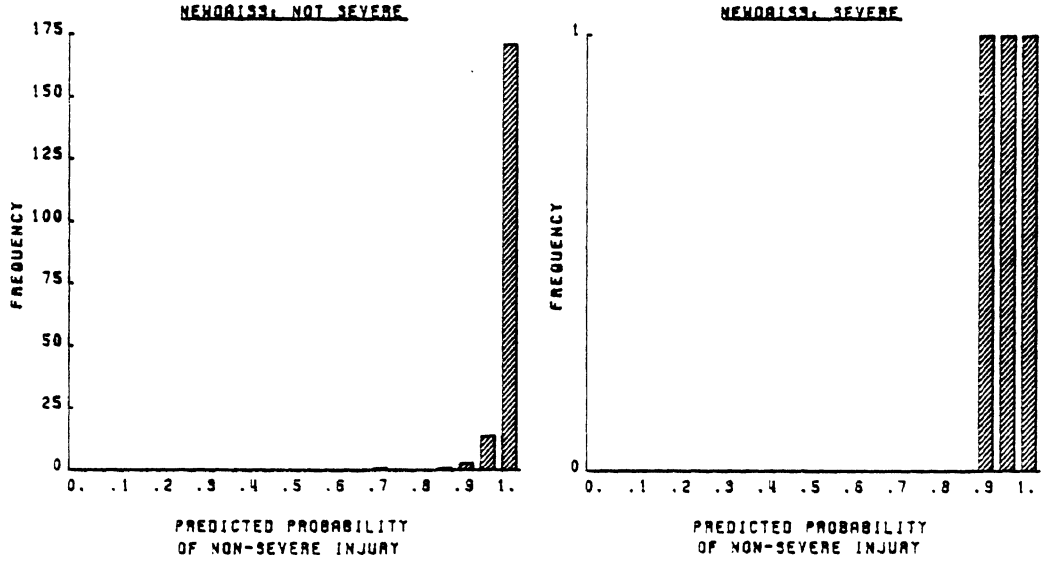

FIGURE 3.8 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For Far NPCD Phase 1 Data - Side Impacts

fixed at 30. The differences in the estimated probabilities of severe injury for the four subsets are clearly shown. For convenience, the probability of severe injury has been plotted. The probability of severe injury is simply 1 - $\hat{p}_i$, where $\hat{p}_i$ is the probability of non-severe injury modelled by Equations 3-20 through 3-23. The figure shows that the occupants of the Near PCD subset had higher probabilities of receiving severe injuries than those of Far PCD, Far NPCD, or Near NPCD, particularly when Lateral Delta V values were less than 35 mph. The

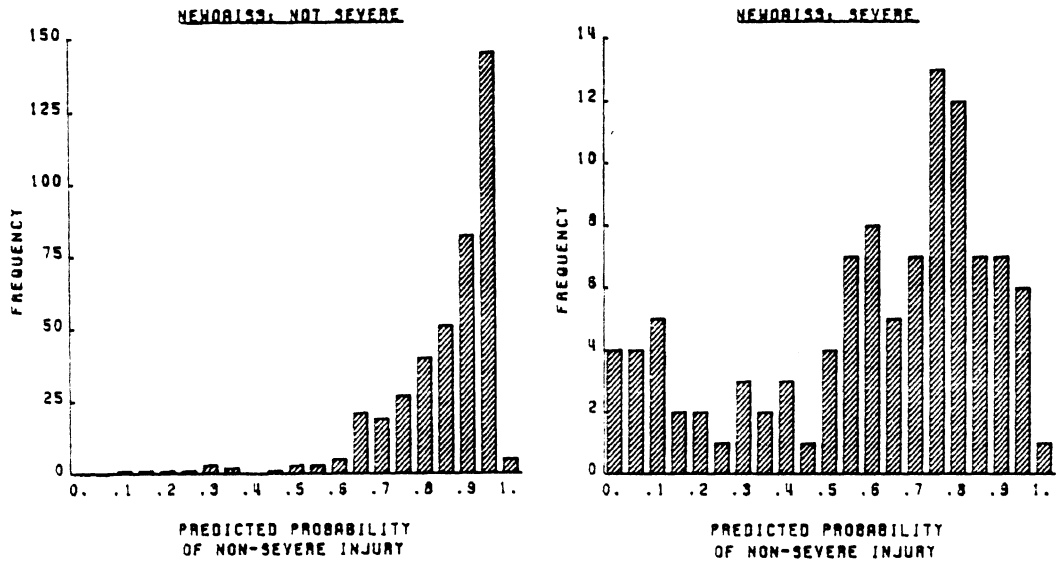

FIGURE 3.9 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For ALL NEAR Phase 1 Data - Side Impacts

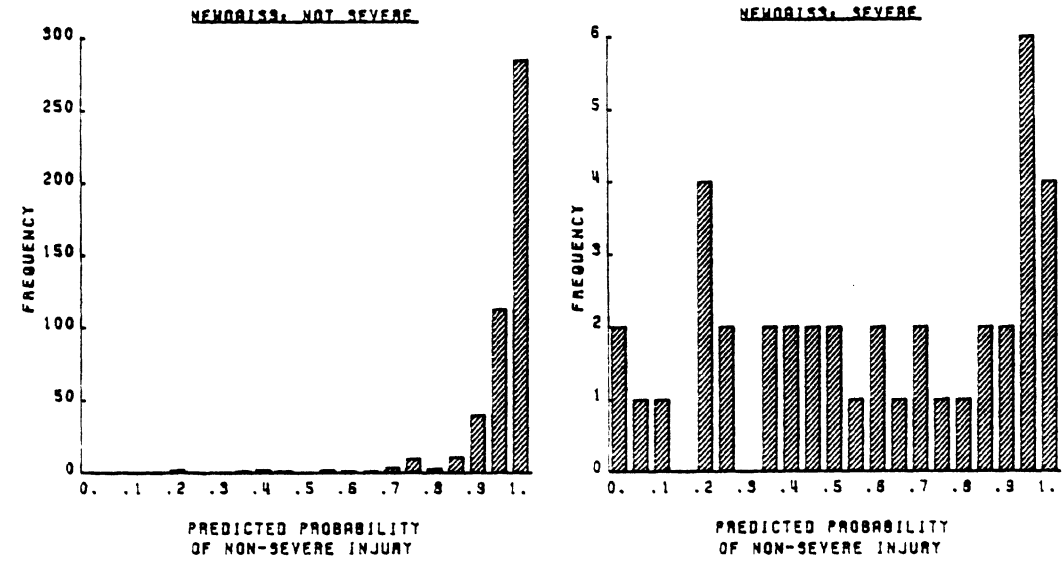

FIGURE 3.10 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For ALL FAR Phase 1 Data - Side Impacts

occupants of the Far PCD subset in turn had higher probabilities of severe injuries than those of the subsets with no damage to the passenger compartments (i.e., Near NPCD and Far NPCD). For Lateral Delta V greater than 30 mph, the probability of a severe injury for an occupant of the Near PCD or Far PCD was close to unity; for an occupant of Near NPCD or Far NPCD to have the probability of a severe injury close to one, Lateral Delta V must be greater than 40 mph. The effect of Age for each subset is shown in Figures 3.12 through 3.15. Each of

these figures has three curves representing Age 20, 40 and 60. The curves show for these ages how the probability of a severe injury varies with Lateral Delta V. In each subset, older occupants have a higher probability of severe injury at any given value of Lateral Delta V. The magnitude of this effect is comparable for all except the Near NPCD group, which shows a substantially larger effect. Figures 3.16 through 3.19 show the 95% confidence intervals for the estimated probability of a severe injury $(1-\hat{p}_i)$ which result from the variance of the coefficients of the independent variables at age 30. Each of these figures, representing a model of a particular subset, consists of three curves. The top curve and the bottom curve indicate the upper bound and the lower bound of the estimated probability of a severe injury respectively while the middle curve is the locus of the estimated probabilities. A narrow band of the confidence limits implies that there is a good chance of reproducing similar modelling results when analysing different sets of data and therefore is a desirable property of a model. The confidence intervals are quite small for the PCD (passenger compartment damage) subsets, but are considerably larger for the NPCD (non-passenger compartment damage) subsets. This is the result of the very small number of cases of the severe injuries (less than 5% of total injuries and less than 10 cases) in each of the latter subsets, which in turn may have made the models less creditable. Finally, Figure 3.20 shows the confidence intervals for both the Near and Far PCD groups on the same graph. The confidence intervals for both subsets do not overlap except when Lateral Delta V becomes quite large at which point the probability of a severe injury approaches one. This implies that the models for the Near PCD subset and the Far PCD subset were in fact different. For Lateral Delta V values of less than 30 mph, near-side occupants are expected to have higher probabilities of receiving severe injuries than far-side passengers when there are damages to the passenger compartment.

Subsampling of Cells. It was noted that there were considerable discrepancies in the number of cases for each level of Lateral Delta V and for each level of Age. That is the distributions of Lateral Delta V and Age, on which the model for each subset was based, were not uniform. To determine if the derived model for each subset was

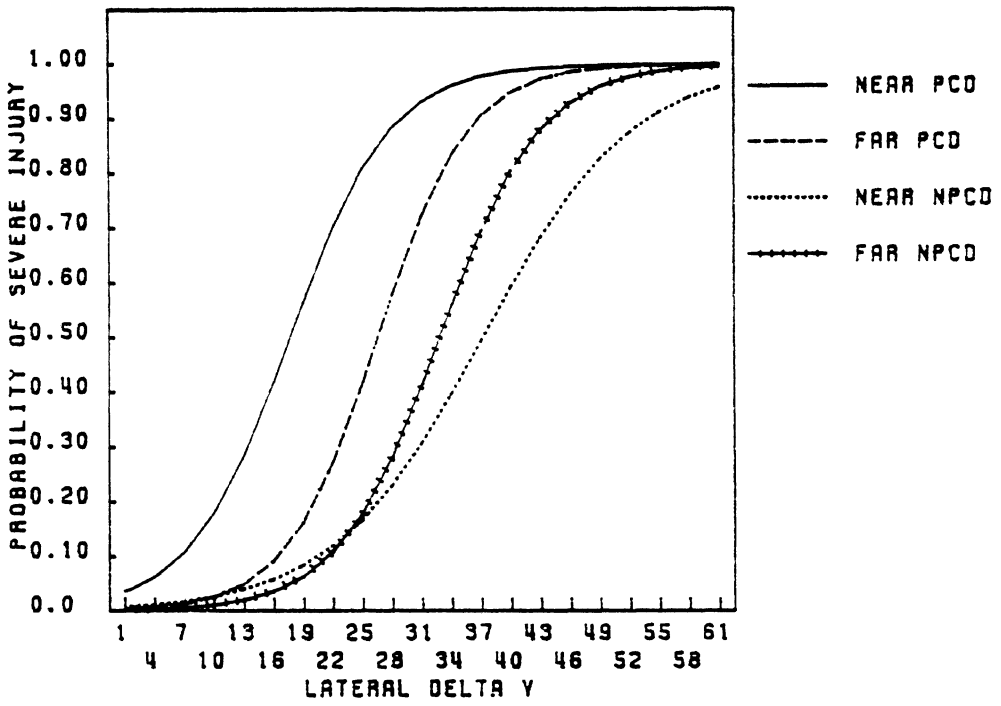

FIGURE 3.11 Logistic Curves of Two-Variable Models (Lateral Delta V, Age) For The Side-Impact Subsets Phase 1 Data - Side Impacts

affected by these uneven cell sizes, subsampling of cells on Lateral Delta V was carried out prior to the actual modelling to yield cells of approximately equal size for each level of Lateral Delta V. The comparison of the univariate models having Lateral Delta V as the independent variable with and without subsampling of the independent variables for the Near PCD subset is shown below:

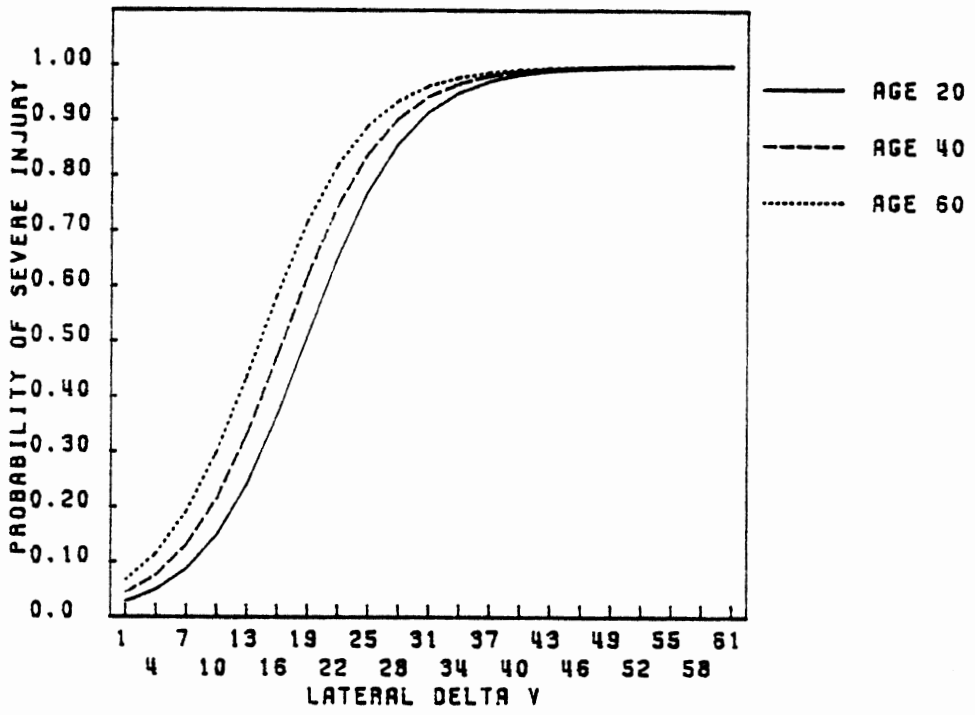

FIGURE 3.12 The Age Effect of Two-Variable Models (Lateral Delta V, Age) For Near PCD Phase 1 Data - Side Impacts

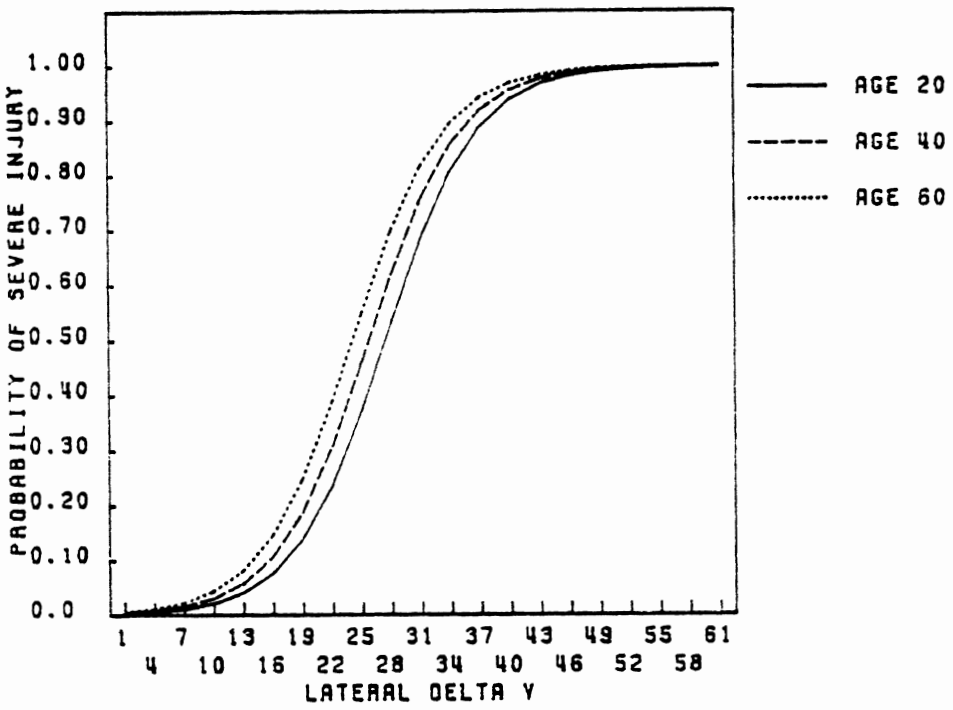

FIGURE 3.13 The Age Effect of Two-Variable Models (Lateral Delta V, Age) For Far PCD Phase 1 Data - Side Impacts

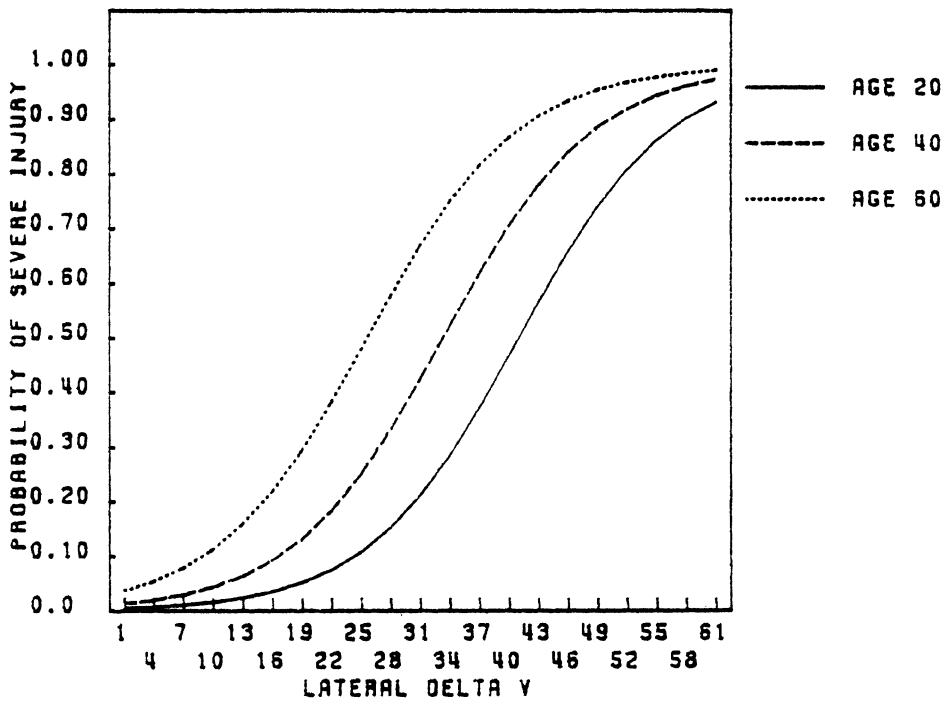

FIGURE 3.14 The Age Effect of Two-Variable Models (Lateral Delta V, Age) For Near NPCD Phase 1 Data - Side Impacts

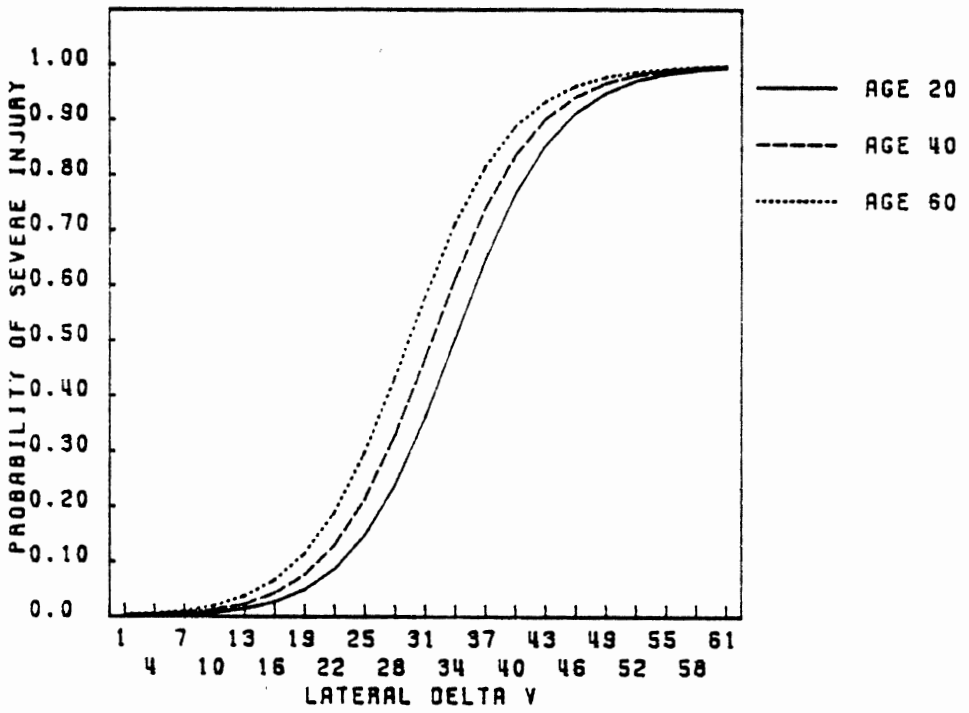

FIGURE 3.15 The Age Effect of Two-Variable Models (Lateral Delta V, Age) For Far NPCD Phase 1 Data - Side Impacts

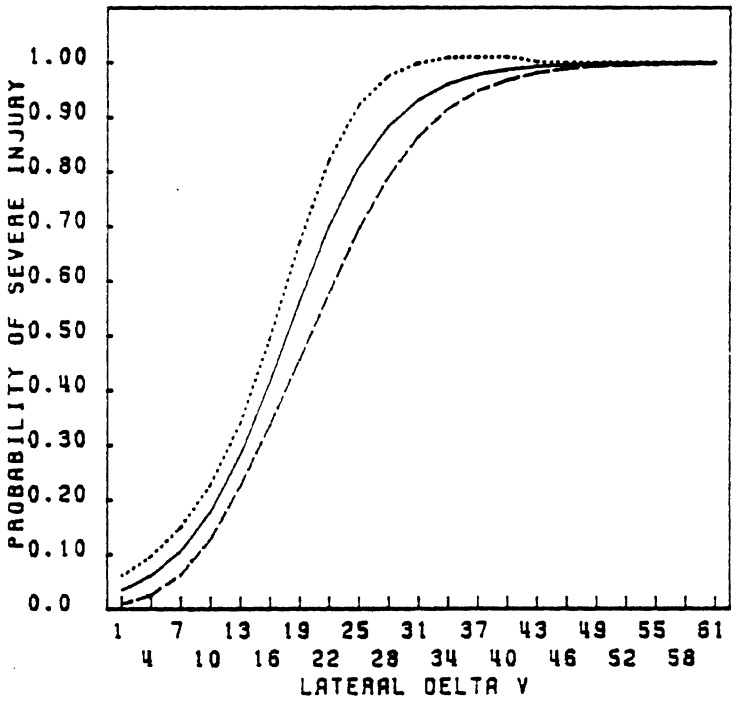

FIGURE 3.16 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Near PCD Phase 1 Data - Side Impacts

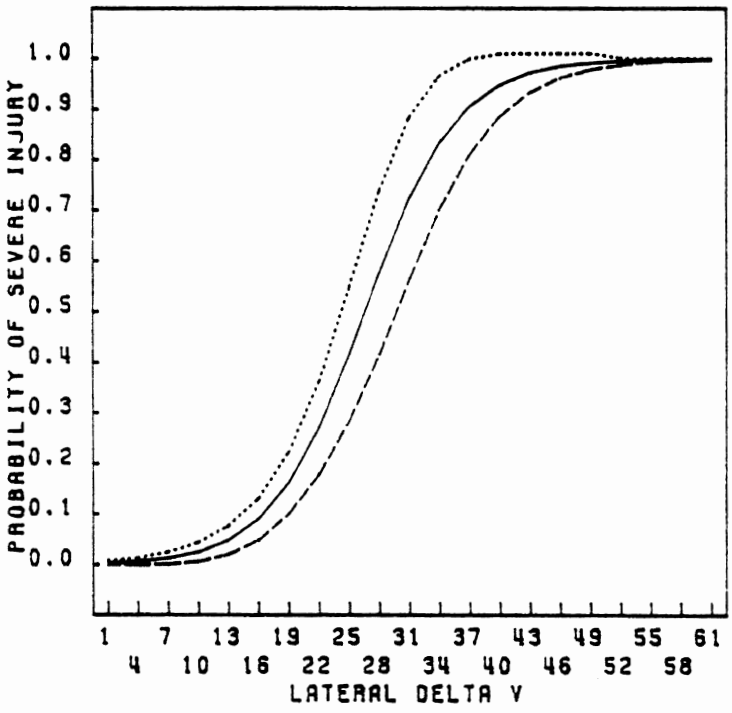

FIGURE 3.17 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Far PCD Phase 1 Data - Side Impacts

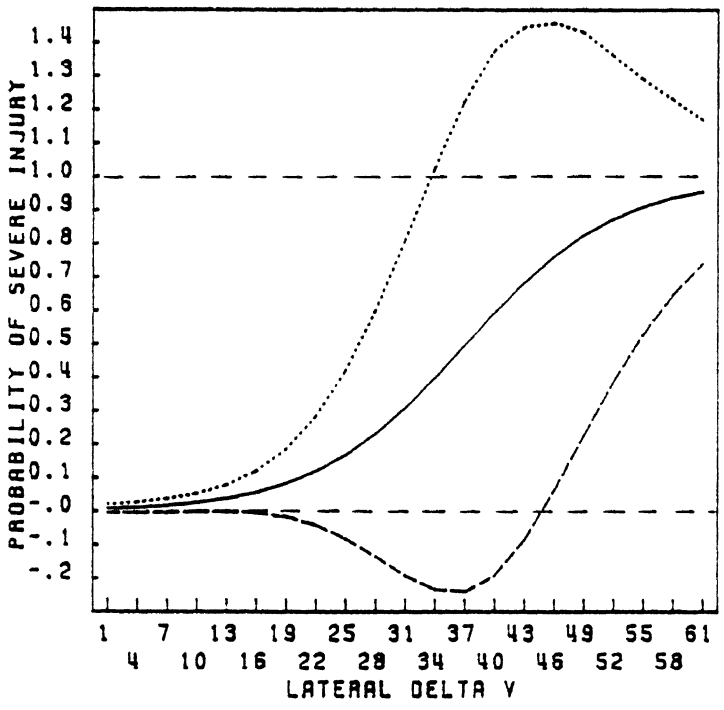

FIGURE 3.18 Confidence Interval of $\hat{p}_j$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Near NPCD Phase 1 Data - Side Impacts

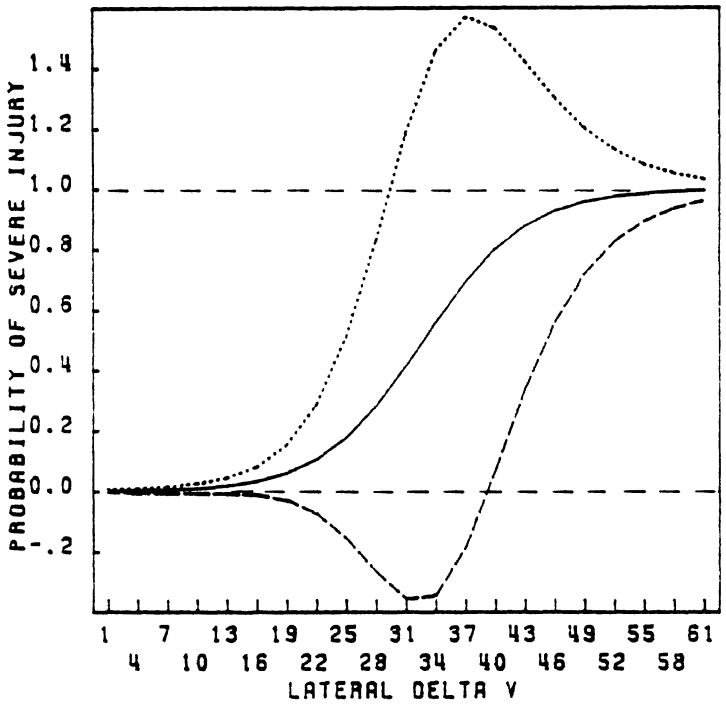

FIGURE 3.19 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Far NPCD Phase 1 Data - Side Impacts

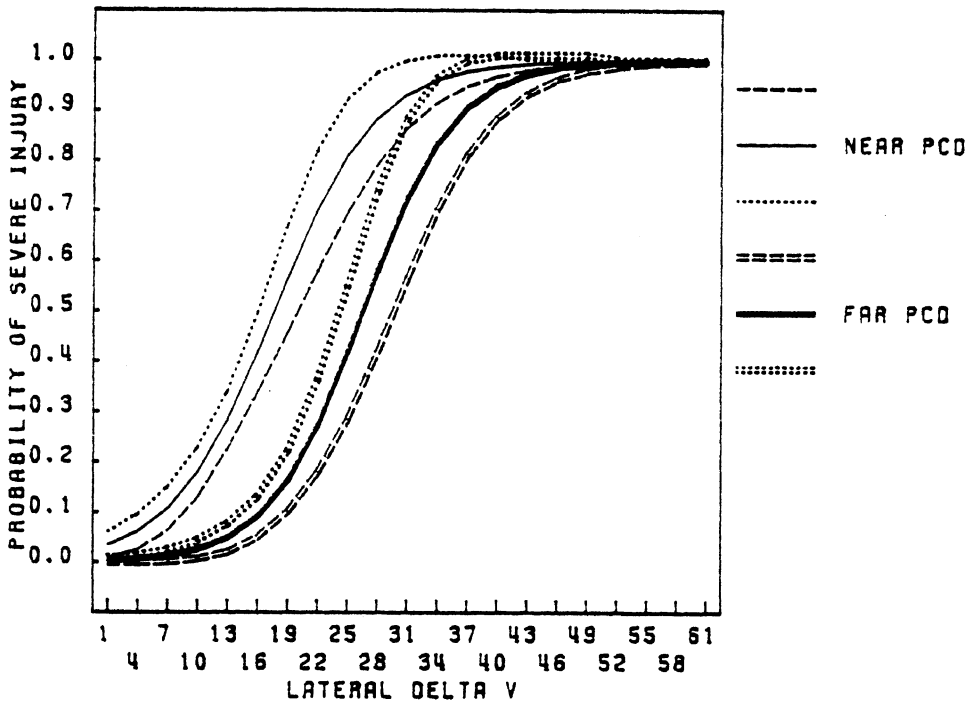

FIGURE 3.20 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Near PCD and Far PCD Phase 1 Data - Side Impacts

TABLE 3.5

Goodness of Fit

Severity = F (Lateral Delta V, Age)

Phase 1 Data - Side Impacts

Subset	Sample Size		Overall % Correct	Non-Severe % Correct	Severe % Correct
	Non-Severe	Severe			
Near PCD	230	95	78	95	36
Far PCD	286	37	93	98	49
Near NPCD	181	9	96	100	11
Far NPCD	190	3	99	100	0
ALL NEAR	411	104	83	97	27
ALL FAR	476	40	95	99	45

Without subsampling (Ratio of severe to non-severe = 0.41)

(3-26) $\hat{p}_i = F(1.802-0.103X)$, LRS=67.71, DF=1, N=326

With subsampling (Ratio of severe to non-severe = 0.54)

(3-27) $\hat{p}_i = F(1.715-0.100X)$, LRS=51.88, DF=1, N=190

The goodness of fit of the two models is shown in Table 3.6.

TABLE 3.6

Goodness of Fit
of Models With and Without Subsampling

Phase 1 Data - Side Impacts

	Sample Size		% Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
Equation (3-26)	231	95	78	95	36
Equation (3-27)	123	67	76	91	48

Note the differences in sample sizes and the ratios of severe to non-severe injuries of both models. The implication of the subsampling of the independent variables, based on the above example, was that while the estimated parameters did not appear to be greatly influenced by the distribution of the independent variable, the percentage correct prediction of severe injuries did improve by 12% when the subsampling was carried out before the model estimation. The difference in sample size (N) of the two models, however, makes it difficult to assess the real merits of the cell subsampling on the model estimation.

3.2.4 Model Evaluation. The modelling results up to now indicated a need for improvement in the prediction of severe injury cases. Note that the severe injuries represent about 14% of the total injuries. Table 3.5 indicated that a model with only Lateral Delta V and Age as the independent variables was capable of predicting the non-severe cases (OAIS 0-2) correctly almost 100% of the time, but the same model

predicted the severe injury cases incorrectly most of the time (about 60%).

To improve the predictive capability of the model, attempts were made to explain those "deviant" cases which had low to moderate Lateral Delta V but OAIS greater than 2. The following analyses were carried out.

For all subsets, the cases were grouped into two classes:

1. Those cases correctly predicted by the model as having low or high injury severity.

2. Those cases incorrectly predicted by the model.

Two-way contingency tables were constructed in which one of the variables was the above "correct/incorrect" variable and the other variable was one of the variables such as Intrusion Location, Ejection, Restraint Usage, Body Region, Injury Type, etc. The number of cases correctly or incorrectly predicted by the model for each level of such variables were therefore known.

Intrusion. A two-way contingency table of the "correct/incorrect" variable and Intrusion Location indicated that the percentage of incorrect prediction of injuries varied with the existence of intrusion and the location of intrusion. For modelling purposes, two analyses were tried.

1. Classification of all cases into intrusion and no-intrusion classes, which were then coded as a dummy variable. A model predicting injury severity was then estimated having Lateral Delta V, Age and this dummy variable as the independent variables. Improvement in the predictive capability of the model was small.

2. Estimation of two separate models for the cases with no intrusion and for the cases with intrusion. The model for cases with no intrusion had only Lateral Delta V and Age as the independent variables. For the model with intrusion, a set of dummy variables was created, based on the levels of the intrusion variable, to reflect intrusion on the sides of a vehicle, on the steering/A-pillar area, on the side override or the combination of these. Improvement in the predictive capability of the resulting models detected was also small.

Ejection. A two-way contingency table of the "correct/incorrect" variable and the degrees of ejection indicated that cases

involving complete ejections tended to have a larger proportion of injury misprediction than those with no-ejection, partial-ejection and entrapment. Incorporation of ejection in the form of a set of dummy variables into the model with Lateral Delta V and Age did not improve the predictive capability of the model.

Restraint Usage. A two-way contingency table of the "correct/incorrect" variable and Restraint Usage indicated that the proportion of injury misprediction for those cases with no restraint installed in the cars were much larger than those with some kinds of restraints available. The model incorporating the Restraint Usage with Lateral Delta V and Age improved the predictive capability more substantially than those incorporating Ejection, but the improvement was small (some 4% of the severe injuries or 3 cases out of 68 cases for the Near PCD subset.)

Restraint Usage and Ejection. A three-way contingency table of the "correct/incorrect" variable, Restraint Usage and Ejection indicated that the proportions of injury misprediction for the various degrees of ejection and entrapment within the no restraint cases did not differ substantially. Incorporating both the Restraint Usage and Ejection into the model did not improve the predictive capability of the model appreciably.

Intrusion and CDC Extent. The CDC Extent was multiplied by the "intrusion/no intrusion" dummy variable to form an interactive variable which described the extent of crush when intrusion was present. This interaction variable, when incorporated into the model with Lateral Delta V and Age, did not improve the model's predictive capability.

Intrusion, CDC Extent and Lateral Delta V. For cases with intrusion, a new variable was created which described the interaction between CDC Extent and Lateral Delta V (CDC Ext. X Lat. Delta V). This new variable, when incorporated into the model with Lateral Delta V and Age, failed to improve the model's predictive capability.

Lateral Delta V, Age and Injury Type. To determine the effects of the injury type on the injury prediction, the Injury Type

variable of Occupant Injury Classification for the first injury (which is the most severe injury) was used in the analysis. The classification of Injury Type in the NCSS file is as follows:

1. laceration
2. contusion
3. abrasion
4. fracture
5. pain
6. concussion
7. hemorrhage
8. avulsion
9. rupture
10. sprain
11. dislocation
12. crushing
13. amputation
14. burn
15. other

The procedure to incorporate the Injury Type variable into the model is as follows:

1. Determine, for each of the six subsets, the number of cases correctly and incorrectly predicted by equations 3-20 through 3-25. Create a variable comprising these two classes and call it a "correct/incorrect" variable.

2. For each subset, construct a two way contingency table of the "correct/incorrect" variable and the Injury Type variable. Calculate the percentage of mispredicted cases within each level of the Injury Type.

3. Isolate those levels of Injury Type which indicated high proportions of mispredicted cases. Create a set of dummy variables of Injury Type corresponding to these isolated levels. Incorporate these dummy variables into the model with Lateral Delta V and Age.

The estimated models for the six subsets are shown below and their goodness of fit is contained in Table 3.7.

Estimated Models with Lateral Delta V, Occupant's Age and Injury Type

Near PCD (N=203, LRS=73.02, DF=5)

(3-28) $\hat{p}_i = F(-0.539 - 0.087X_1 - 0.007X_2 + 1.677X_3 - 2.161X_4 - 2.146X_5)$

Far PCD (N=168, LRS=57.81, DF=5)

(3-29) $\hat{p}_i = F(1.982 - 0.102X_1 - 0.002X_2 - 0.100X_3 - 0.777X_5 - 0.467X_6)$

Near NPCD (N=61, LRS=22.62, DF=5)

(3-30) $\hat{p}_i = F(0.659 + 0.050X_1 - 0.024X_2 + 0.137X_3 - 3.560X_5 + 0.176X_7)$

Far NPCD (N=54, LRS=7.72, DF=4)

(3-31) $\hat{p}_i = F(4.613 - 0.051X_1 - 0.012X_2 - 2.859X_3 - 2.040X_8)$

All Near (N=264, LRS=78.55, DF=4)

(3-32) $\hat{p}_i = F(-0.572 - 0.074X_1 - 0.0049X_2 + 1.539X_3 - 4.050X_5)$

All Far (N=222, LRS=57.56, DF=3)

(3-33) $\hat{p}_i = F(2.482 - 0.090X_1 - 0.005X_2 - 0.615X_3)$

$\hat{p}_i$ is the estimated probability of a non-severe injury,
F is the logistic function,
X_1 is Lateral Delta V,
X_2 is Age,
X_3 is 1 if injury is a Fracture, -1 otherwise,
X_4 is 1 if injury is a rupture, -1 otherwise,
X_5 is 1 if injury is a Dislocation, -1 otherwise,
X_6 is 1 if injury is a Sprain, -1 otherwise,
X_7 is 1 if injury is a Laceration, -1 otherwise,
X_8 is 1 if injury is a Contusion, -1 otherwise, and
LRS is the Likelihood Ratio Statistic.

TABLE 3.7

Goodness of Fit

Severity = F(Lateral Delta V, Age, Injury-Type)

Phase 1 Data - Side Impacts

Subset	Sample Size		% Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
Near PCD	122	81	72	80	61
Far PCD	141	27	88	95	48
Near NPCD	54	7	90	100	14
Far NPCD	51	3	94	100	0
ALL NEAR	176	88	74	83	55
ALL FAR	192	30	88	95	37

Note that the Injury Type variable incorporated into the models of the four subsets as dummy variables differed from one subset to another, as follows:

Near PCD The dummy variables were:
- fracture
- rupture
- dislocation

Far PCD The dummy variables were:
- fracture
- sprain
- dislocation

Near NPCD The dummy variables were:
- laceration
- fracture
- dislocation

Far NPCD The dummy variables were:
- contusion
- fracture

All Near The dummy variables were:
- fracture
- dislocation

All Far The dummy variable was fracture

Lateral Delta V, Age and Body Region. Similar procedures were used in bringing Body Region into the model. The model estimation

results of the six subsets are shown below and their goodness of fit are contained in Table 3.8.

TABLE 3.8

Goodness of Fit
Severity = F(Lateral Delta V, Age and Body Region)

Phase 1 Data - Side Impacts

Subset	Sample Size		% Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
Near PCD	123	83	75	92	49
Far PCD	140	28	89	96	54
Near NPCD*	54	7	90	96	43
Far NPCD	51	3	94	100	0
ALL NEAR*	177	90	77	87	54
ALL FAR	191	31	91	98	42

*Age was found to be not significant

The body region variable incorporated into the models of the four subsets as dummy variables differed from one subset to another, as shown below:

Near PCD The dummy variables were:
- forearm
- abdomen
- lower extremities
- ankle-foot

Far PCD The dummy variables were:
- shoulder
- chest
- abdomen
- pelvic-hip

Near NPCD The dummy variables were:
- shoulder
- elbow
- chest
- abdomen

Far NPCD The dummy variables were:
- face
- neck

chest

All Near The dummy variables were:
- shoulder
- elbow/forearm
- chest
- abdomen
- lower extremities, leg, ankle/foot

All Far The dummy variables were:
- shoulder
- chest
- pelvic/hip

Equations 3-28, 3-29, 3-31, 3-32, and 3-33 for Near PCD, Far PCD, Far NPCD, All Near, and All Far each appeared reasonable in terms of the signs of the coefficients, that is the probability of a non-severe injury increases as Lateral Delta V decreases and as Age decreases. Equation 3-30 (Near NPCD) appeared to be an unreasonable model because of the positive sign of the coefficient of Lateral Delta V, which countered intuition. The Injury Type variable warrants some attention here. The signs of all dummy variables in Equations 3-28 through 3-33, with the exception of that of fracture, were negative as anticipated. The fracture dummy variable, however, had negative coefficients for far-sided passengers and positive coefficients for near-sided passengers. This may be explained by the fact that fracture injuries could receive OAIS of either 0-2 or 3-6, and that the overall number of fracture injuries involve many body regions; such characteristics are generally not true with the other injury types such as rupture and dislocation, which are almost exclusively causes of OAIS 3-6. In predicting the occurrence of a severe injury , fracture is a weaker explanatory variable of severe injuries than, say, rupture or dislocation. The models represented by Equations 3-28 to 3-33 showed mixed results as far as the improvement in the models' predictive capability was concerned. For Near PCD, the model (Equation 3-28) now predicted 61% of the severe injury cases correctly, (an improvement of 25% for severe injuries) while it now only predicted 80% of the non-severe injury cases correctly, (a reduction of 15% of total non-severe injury cases). For Far PCD, no improvement in prediction was accomplished by the new model (Equation 3-29). For Near PCD, the improvement accomplished was negligible. For Far NPCD, no improvement was accomplished. For All Near, the gain in percentage correct prediction of the severe injury

cases was 28% while the reduction in percentage correct prediction of the non-severe injury cases was 14%. For All Far, there was no improvement gained with Injury Type. Note that the models for the two subsets with no damage to passenger compartments (Near NPCD and Far NPCD) are likely to be less stable than those for other subsets due to their very small numbers of severe injuries.

Estimated Models with Lateral Delta V, Age and Body Region

Near PCD (N=206, LRS=56.44, DF=6)

(3-34) $\hat{p}_i = F(0.207 - 0.075X_1 - 0.014X_2 + 0.533X_{11} - 0.305X_{12} - 3.064X_{13} + 1.181X_{14})^1$

Far PCD (N=168, LRS=54.65, DF=6)

(3-35) $\hat{p}_i = F(2.071 - 0.083X_1 - 0.004X_2 - 0.165X_{12} + 0.286X_{15} - 0.317X_{16} - 0.784X_{17})$

Near NPCD* (N=61, LRS=21.63, DF=5)

(3-36) $\hat{p}_i = F(3.783 - 0.192X_1 - 1.865X_{12} - 1.335X_{16} - 1.951X_{15} - 1.667X_{18})$

Far NPCD (N=54, LRS=7.26, DF=5)

(3-37) $\hat{p}_i = F(3.145 - 0.053X_1 - 0.002X_2 - 1.392X_{16} - 1.135X_{19} - 1.798X_{20})$

All Near* (N=267, LRS=78.90, DF=6)

(3-38) $\hat{p}_i = F(0.971 - 0.0697X_1 + 0.555X_{15} + 0.137X_{21} - 0.356X_{16} - 1.161X_{12} + 0.067X_{22})$

All Far (N=222, LRS=56.38, DF=5)

(3-39) $\hat{p}_i = F(2.220 - 0.089X_1 - 0.004X_2 + 0.381X_{15} - 0.309X_{16} - 0.862X_{17})$

where

X_1 is Lateral Delta V,
X_2 is Age,
X_{11} is 1 if it is Forearm, -1 otherwise,
X_{12} is 1 if it is Abdomen, -1 otherwise,
X_{13} is 1 if it is Lower Extremities, -1 otherwise,
X_{14} is 1 if it is ankle/foot, -1 otherwise,
X_{15} is 1 if it is Shoulder, -1 otherwise,
X_{16} is 1 if it is Chest, -1 otherwise,
X_{17} is 1 if it is Pelvis/Hip, -1 otherwise,
X_{18} is 1 if it is Elbow, -1 otherwise,
X_{19} is 1 if it is Face, -1 otherwise,
X_{20} is 1 if it is Neck, -1 otherwise,
X_{21} is 1 if it is Elbow/Forearm, -1 otherwise, and
X_{22} is 1 if it is Lower Extremities, Leg, Ankle/Foot, -1 otherwise.

*Age was not significant

The models having Lateral Delta V, Age and Body Region as the independent variables for the six subsets as shown in Equations 3-34 to 3-39 appeared reasonable in terms of the signs of the coefficients of Lateral Delta V and Age. For Near NPCD and All Near, Age was found to be not significant. The signs of the coefficients of the Body Region dummy variables showed consistency--abdomen, chest, pelvis/hip, always had negative coefficients, while ankle/foot, leg always had positive coefficients with the exception of Near NPCD; elbow/forearm always had positive coefficients. The implication of this was that an injury to Abdomen or Chest or Pelvic/Hip is more likely to receive a severe injury than an injury to Ankle/Foot or Leg or Elbow or Forearm. This in turn was confirmed by the two-way contingency table of NEWOAIS3 and Body Region, which indicated that the proportions of a severe injury to total injuries were much higher for Abdomen and Chest than for Forearm, Elbow, Ankle/Foot and Leg.

The models represented by Equations 3-34 to 3-39 showed mixed results in terms of their improved predictive capability relative to the models with only two continuous variables (Delta V and Age). For Near PCD, the gain in percentage correct prediction of the severe injury cases was 13% while the reduction in percentage correct prediction of the non-severe injury cases was 3%. For Far PCD, the gain in percentage correct prediction of the severe injury cases was 5% while the reduction in percentage correct prediction of the non-severe injury cases was 2%. For Near NPCD, the gain was 32% for severe injuries and the reduction for non-severe injuries was 4%. For Far NPCD, no improvement resulted. For All Near the gain was 27% for severe injuries and the reduction was 10% for cases with non-severe injuries; and for All Far, there was no improvement. Again, the models for Near NPCD and Far NPCD are likely to be less stable than those for other subsets due to their very small numbers of cases of severe injuries.

Assessment of the Effects of Injury Type and Body Region.

The goodness of fit of the models represented by Equations 3-28 to 3-39 seemed to indicate that while Injury Type and Body Region appeared to improve the models' predictive capability by better predicting the severe injuries for near passengers, they did not do very much for the

models involving far-side passengers. For Near PCD, the Injury Type variable strongly influenced the prediction of severe injury cases because:

1. Rupture and dislocation injuries in this subset almost immediately implied NEWOAIS3 severe.

2. There were altogether 4 rupture injuries in Near PCD with valid supplementary information, three of which had been mispredicted by the two-variable model (i.e., model with Lateral Delta V and Age); there were 6 dislocation injuries, four of which had been mispredicted.

3. There were altogether 68 fracture injuries in Near PCD, 38 of which had been mispredicted. The probability of a severe injury with fracture is about two thirds. Unlike rupture and dislocation, the proportion of severe fracture injuries varied with different body regions. For example, all fractured forearms were associated with NEWOAIS3 severe, so were fractured lower extremities, ankle-foot and neck; 70% of the fractured faces also received NEWOAIS3 severe, and 65% of fractured chest got NEWOAIS3 severe, etc.

The implication is that for Near PCD a Rupture, a Dislocation, or a Fracture is prone to high injury severity, quite frequently regardless of the associated values of Lateral Delta V. The mispredicted cases by the two-variable model were frequently the severe injuries with low to moderate values of Lateral Delta V.

For Far PCD, 10 cases out of the 24 fracture injuries had been mispredicted, one out of the two dislocation injuries had been mispredicted and 1 out of the 3 sprains had been mispredicted. These three Injury Types could receive either NEWOAIS3 severe or non-severe depending on the affected body regions. It appears that for a far-side passenger to have received a severe injury the Lateral Delta V involved would have had to be quite high. And therefore were likely to be readily correctly predicted by the two-variable models (Lateral Delta V and Age). Although there were some severe fractures, dislocations, and sprains which had low Lateral Delta V and therefore mispredicted by the two-variable model, such cases occurred less frequently than in the cases of near-side passengers. As a result, the injury-type variable did not help to predict severe injuries as well for Far PCD as it did for Near PCD. In fact for Far PCD the body region variable appeared to be a better prediction variable than Injury Type.

3.2.5 Final Models. For each subset, the dummy variables for Injury Type and Body Region that had been created earlier were added to Lateral Delta V and Age. The estimated models for the six subsets are shown below. The estimated models for Near NPCD and Far NPCD appeared unstable, the models did not converge and the constant term and the coefficients of the logit function were extremely large. This was due to their very small numbers of severe injuries.

Estimated Models with Lateral Delta V, Age, Injury Type and Body Region

Near PCD (N=203, LRS=104.12, DF=9)

(3-41) $\hat{p}_i = F(-1.412 - 0.082X_1 - 0.012X_2 + 0.569X_{11} - 0.638X_{12} - 2.608X_{13} + 1.090X_{22} + 1.192X_3 - 0.653X_4 - 2.859X_5)$

Far PCD (N=167, LRS=72.41, DF=9)

(3-42) $\hat{p}_i = F(1.414 - 0.094X_1 - 0.0005X_2 - 0.110X_3 - 0.452X_5 - 0.757X_6 + 0.612X_{15} - 0.281X_{16} - 0.703X_{12} - 0.522X_{17})$

Near NPCD (N = 61)
Unstable**

Far NPCD (N = 54)
Unstable**

All Near* (N=264, LRS=140.86, DF=8)

(3-43) $\hat{p}_i = F(-0.887 - 0.073X_1 + 0.823X_{15} + 0.513X_{21} - 0.379X_{16} - 1.908X_{12} + 0.364X_{22} + 1.374X_3 - 4.026X_5)$

All Far (N=221, LRS=70.45, DF=6)

(3-44) $\hat{p}_i = F(1.901 - 0.088X_1 - 0.003X_2 - 0.586X_3 + 0.582X_{15} - 0.192X_{16} - 1.176X_{17})$

where $X_1, X_2, \ldots, X_{22}$ are similarly defined as those in Equations 3-28 through 3-33

*Age was not significant.
**Models did not converge.

Goodness of fit of the models are shown in Table 3.9. Considerable improvement in the correct prediction of high-severity injuries were noted for all subsets. These models predicted the high severity cases

better than when the models had only either Injury Type or Body Region in addition to Lateral Delta V and Age

Figures 3.21 and 3.22 show the histograms of $\hat{p}_i$ for the Near PCD and the Far PCD subsets. Each of these figures consists of two histograms, one for non-severe injuries and the other for severe injuries. Both histograms have the same axes, one designating the $\hat{p}_i$ values at the 0.05 interval and the other the number of cases with the particular $\hat{p}_i$ values. Comparisons of Figure 3.21 with Figure 3.5 and Figure 3.22 with Figure 3.6 indicate that the additional independent variables, namely, Injury Type and Body Region improved the models predictive capability for severe injuries considerably for both subsets, particularly for Near PCD.

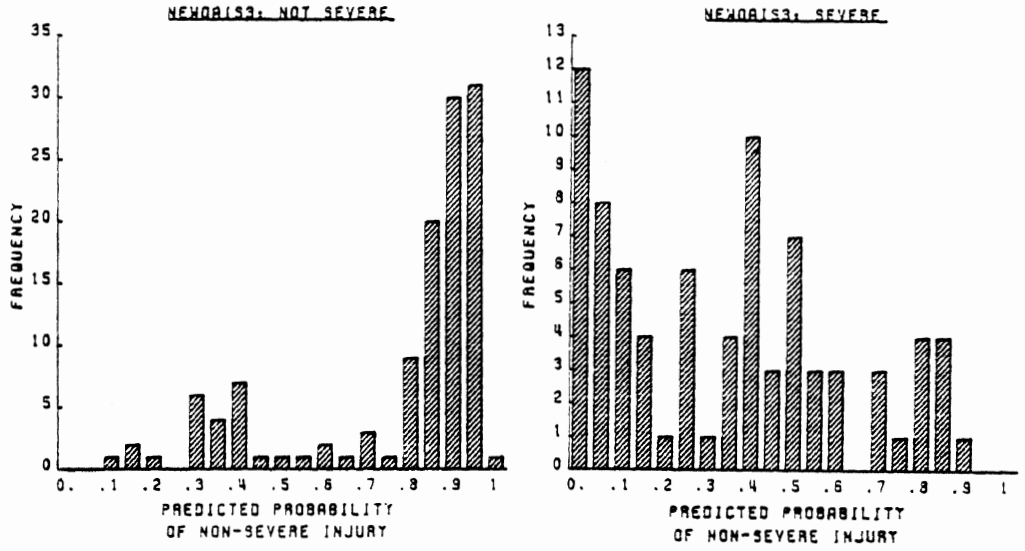

FIGURE 3.21 Histograms of $\hat{p}_i$ of Four-Variable Model (Lateral Delta V, Age, Body Region and Injury Type) For Near PCD Phase 1 Data - Side Impacts

The models for All Near (Equation 3-43) indicated that Age was not significant. The difference in the models for near-side and far-side passengers emerged somewhat more clearly from these models (Equations 3-41 to 3-44) as follows:

1. Fracture injuries for near-side passengers always had positive coefficients whereas those for far-side passengers always had negative coefficients.

2. The constant terms for near-side passenger were negative quantities while those for far-side passenger were positive.

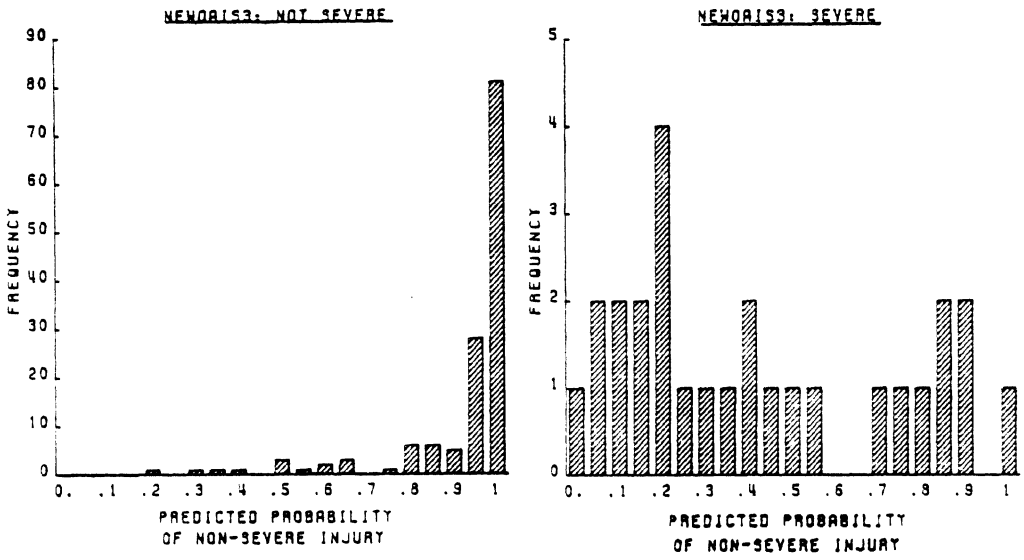

FIGURE 3.22 Histograms of $\hat{p}_i$ of Four-Variable Model (Lateral Delta V, Age, Body Region and Injury Type) For Far PCD Phase 1 Data - Side Impacts

TABLE 3.9

Goodness of Fit

Severity = F(Lateral Delta V, Age, Body Region, Injury Type)

Phase 1 Data - Side Impacts

Subset	Sample Size		% Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
Near PCD	122	31	78	82	73
Far PCD	140	27	90	95	63
Near NPCD	54	7	unstable	unstable	unstable
Far NPCD	51	3	unstable	unstable	unstable
ALL NEAR*	176	88	83	89	72
ALL FAR	191	30	91	97	53

*Age was redundant

3. The two-way contingency table between NEWOAIS3 and Injury Type indicated that the chance of fracture injuries to be severe was about 62% for near-side passengers and about 46% for far-side passengers.

4. It was also noted that for far-side passengers, both the injury type and the body region, when incorporated into the models, caused only a small reduction in the correct prediction of the low severity cases.

Near-side and far-side passengers displayed the following similarities:

1. Coefficients for Lateral Delta V and Age of both were comparable.

2. Dislocation injuries, chest and abdomen always had negative coefficients. The two-way contingency table between NEWOAIS3 and dislocation indicated that dislocation resulted in severe injuries to near-side passengers more often than far-side passengers. Injuries to the chest and abdomen were more likely to be severe for near-side passengers than for far-side passengers.

By adding Injury Type and Body Region into the models, the sample size (N) in the modelling of each subset was reduced considerably due to the missing data of the injury-type and the body-region being excluded from the analysis. Table 3.10 shows the number of valid cases for each subset. Table 3.11 shows the missing cases for Near PCD and Far PCD by OAIS by the Lateral Delta V values.

For Near PCD, the missing cases should not significantly affect the modeling results. From the results of the earliest models (i.e., NEWOAIS3 as a function of Lateral Delta V and Age) it was recognized that the mispredictions of injuries occurred in the cases with low to moderate Lateral Delta V but NEWOAIS3 severe. The absence of many of such cases could discredit the modelling results when Injury Type and Body Region entered the models. The absence of cases with low OAIS (0-2) should not affect the subsequent modelling results in any way because they were the cases which the earlier models could predict quite accurately.

For Far PCD, however, by bringing Injury Type and Body Region into the model, ten cases of injuries with low to moderate Lateral Delta V and NEWOAIS3 severe became lost. This represented 50% of the originally mispredicted cases which would have needed further investigation and analyses.

TABLE 3.10

Comparison of The Number of Cases For
Models With and Without
The Dummy Variables

Phase 1 Data - Side Impacts

Subset	Independent Variables Chosen for the Model	Sample Size	
		Non-Severe	Severe
Near PCD	Lateral Delta V, Age	230	95
	Lateral Delta V, Age, Injury Type, Body Region	122	81
Far PCD	Lateral Delta V, Age	286	37
	Lateral Delta V, Age, Injury Type, Body Region	140	27
Near NPCD	Lateral Delta V, Age	181	9
	Lateral Delta V, Age, Injury Type, Body Region	54	7
Far NPCD	Lateral Delta V, Age	190	3
	Lateral Delta V, Age, Injury Type, Body Region	51	3
ALL NEAR	Lateral Delta V, Age	411	98
	Lateral Delta V, Injury Type, Body Region	176	88
ALL FAR	Lateral Delta V, Age	476	41
	Lateral Delta V, Age, Injury Type, Body Region	191	30

TABLE 3.11

Missing Cases by NEWOAIS3 Codes and Lateral Delta V

Phase 1 Data - Side Impacts

Subset	No. of cases mispredicted by earlier model		No. of Missing Cases			
			Non-Severe		Severe	
	Non-Severe	Severe	Lateral Delta V (0-20 mph)	Lateral Delta V (20+ mph)	Lateral Delta V (0-20 mph)	Lateral Delta V (20+ mph)
Near PCD	11	61	107	1	7	7
Far PCD	5	19	146	0	10	0

It is worth noting that the estimated models for different subsets, which have Lateral Delta V, Age, Injury Type and Body Region as the independent variables, all displayed one common characteristic. The coefficients of the injury-type dummy variables and the body-region dummy variables were considerably larger in magnitude than those of the Lateral Delta V variable and Age. This seems to imply that given a side impact accident the Injury Type and the affected body region combined can, especially for the severe cases, almost predict the resulting injury severity without information on crash severity. A cursory examination of the three-way tables of NEWOAIS3, Injury Type and Body Region confirmed this. Injury Type and Body Region are felt to be bad choices for the independent variables because the combination of both immediately reflects, for certain injury types and certain body regions, the NEWOAIS3 coding. From the ideal model viewpoint the models should use crash severity to predict injury severity, injury type and affected body region. This leads to the question of whether Injury Type and Body Region should be selected as the independent variables or whether other more "causal" variables should be found to replace these two variables. Intuitively, Lateral Delta V together with Contact Point and the position of occupants relative to the impact would be expected to give some information about the types of injuries suffered by particular body regions. If such were the case, Contact Point might prove to be a more desirable independent variable than both the Injury Type and Body Region variables.

The development of mechanistic models for side impacts is continued with the addition of the Phase 2 data in the next section.

3.3 Final Analytical Results for Side Impacts

The mechanistic model development was continued when the Phase 2 data files were prepared. Perhaps the most critical task was the validation of the Phase 1 models with the Phase 2 data as discussed in Section 2. One would hope that the relationships developed with the Phase 1 data would reflect the basic physical principles governing the collision event, and, therefore, would be stable. Since Phase 2 is simply a continuation of NCSS, it would indeed be disappointing if the results of the modelling efforts were appreciably different for the two Phases. The Phase 1 models are validated by determining their predictive capability when applied to the Phase 2 data. The Phase 2 data is also used to re-estimate the coefficients of the Delta V and Age models. These results form a basis for the combination of the Phase 1 and the Phase 2 data as described in Section 3.3.3. With the increased sample size through combining the data of both phases, additional variables are again considered for the models. Efforts are made to combine the subsets also. Contact Point is reviewed, and the final models are presented in Section 3.3.5. This section closes with a discussion of the significant findings.

3.3.1 Validation of the Phase 1 Models. The estimated Phase 1 models as represented by Equations 3-20 to 3-25 were applied to the Phase 2 data and a goodness of fit for the six subsets was obtained as shown in Table 3.12.

Comparison of Table 3.12 with the goodness of fit of the Phase 1 data (Table 3.5) reveals that:

1. For each subset, the overall proportion of cases correctly predicted by the models for the Phase 2 data was 5% to 8% lower than for the Phase 1 data.

2. For Near PCD and All Near, the models actually predicted the severe injury cases about 6% to 9% better for the Phase 2 data than for the Phase 1 data. However, for these same subsets the proportion of the correct prediction for non-severe injuries was about 3% to 7% lower for the Phase 2 data than for the Phase 1 data.

TABLE 3.12

Goodness of Fit

Severity = F(Lateral Delta V, Age)

Phase 2 Data - Side Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
Near PCD	174	91	73.2	87.9	45.1
Far PCD	218	41	86.9	98.2	26.8
Near NPCD	106	10	91.4	100.0	0
Far NPCD	96	9	91.4	100.0	0
ALL NEAR	280	101	77.4	93.6	32.7
ALL FAR	314	50	88.2	99.0	20.0

3. For Far PCD and All Far, the models predicted the severe injury cases in the Phase 2 data only half as well as they did in the Phase 1 data. The prediction of non-severe injuries for these subsets, however, were identical in the Phase 2 data and the Phase 1 data.

TABLE 3.13

Lateral Delta V

Phase 2 Data - Side Impacts

Subset	Sample Size	Range	Mean	S.D.
Near PCD	330	2-42	13.14	6.93
Far PCD	309	2-49	13.27	6.93
Near NPCD	136	2-24	8.64	4.05
Far NPCD	127	2-22	9.27	4.13
All Near	465	2-42	11.83	6.14
All Far	436	2-49	12.10	6.50

3.3.2 Model Estimation - Phase 2 Data. Table 3.13 gives the details on Lateral Delta V in the Phase 2 data for the six subsets, while Table 3.14 gives the proportions of severe and non-severe injuries for the six subsets. The ranges of Lateral Delta V for near-side occupants and far-side occupants were comparable. The ranges for cases with passenger compartment damage, however, were much larger than for cases with non-passenger compartment damage. Near-side occupants in vehicles with passenger compartment damage were far more likely to sustain high severity injuries than either far-side occupants or occupants of vehicles with non-passenger compartment damage. The ranges of Lateral Delta V and the proportions of severe injuries to total injuries of the subsets in the Phase 2 data were comparable with those in the Phase 1 data.

TABLE 3.14

Injury Proportion

Phase 2 Data - Side Impacts

Subset	Sample Size	Percent of Non-Severe Injuries	Percent of Severe Injuries
Near PCD	445	67.6	32.4
Far PCD	421	85.5	14.5
Near NPCD	194	93.8	6.2
Far NPCD	176	93.2	6.8
All Near	639	75.6	24.4
All Far	597	87.8	12.2

The goodness of fit results contained in Table 3.12 were based on fitting the Phase 1 data models to the Phase 2 data. It is envision that still better goodness of fit would be obtained if the estimation was based on the Phase 2 data. Since Lateral Delta V and Age appeared to be significant explanatory variables of injury severity in the Phase 1 models, the coefficients of these two independent variables therefore should be estimated using the Phase 2 data. The goodness of fit of these new models should then be determined and compared with that contained in Table 3.5 and Table 3.12.

The model estimation results of the Phase 2 data with Lateral Delta V and Age as the independent variables are as follows:

Estimated Models with Lateral Delta V and Age

Near PCD (N=265, LRS=52.19, DF=2)

(3-45) $\hat{p}_i = F(2.0218 - 0.0781X_1 - 0.0166X_2)$

Far PCD (N=259, LRS=46.62, DF=2)

(3-46) $\hat{p}_i = F(2.6633 - 0.0897X_1 - 0.0097X_2)$

Near NPCD (N=116, LRS=19.59, DF=2)

(3-47) $\hat{p}_i = F(4.2348 - 0.1509X_1 - 0.0301X_2)$

Far NPCD (N=105, LRS=19.12, DF=2)

(3-48) $\hat{p}_i = F(3.5252 - 0.1844X_1 - 0.0019X_2)$

All Near (N=381, LRS=86.39, DF=2)

(3-49) $\hat{p}_i = F(2.4584 - 0.0975X_1 - 0.0174X_2)$

All Far (N=364, LRS=65.48, DF=2)

(3-50) $\hat{p}_i = F(2.7960 - 0.0988X_1 - 0.0095X_2)$

where

$\hat{p}_i$ is the probability of a non-severe injury,
F is the logistic distribution,
X_1 is Lateral Delta V,
X_2 is Age, and
LRS is the Likelihood Ratio Statistic.

The goodness of fit of the models represented by Equations 3-45 to 3-50 is shown in Table 3.15.

Comparison of the goodness of fit results of the Phase 2 data 2-variable models (Lateral Delta V and Age) as shown in Table 3.15 with those of the Phase 1 data 2-variable models (Table 3.5) reveals that:

1. The percent overall correct prediction of the Phase 2 data models was lower than that of the Phase 1 data models, about

TABLE 3.15

Goodness of Fit

Injury Severity = F(Lateral Delta V, Age)

Phase 2 Data - Side Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
Near PCD	174	91	75.1	90.8	45.1
Far PCD	218	41	86.5	97.7	26.8
Near NPCD	106	10	92.2	99.1	20.0
Far NPCD	96	9	92.4	97.9	33.3
All Near	280	101	80.1	94.6	39.6
All Far	314	50	87.9	98.4	22.0

3-4% lower for near-side occupants, and about 6-7% for far-side occupants.

2. For near-side occupants, the percent correct prediction of severe injuries of the Phase 2 models was some 10% higher than that of the Phase 1 models, the percent correct prediction for non-severe injuries of the Phase 2 data models was about 4% lower. For far-side occupants, the Phase 2 models predicted severe injuries absent 20% worse than the Phase 1 models but they predicted non-severe injuries just as well.

Comparison of the goodness of fit measures for the Phase 2 models in Table 3.15 and the goodness of fit measures of the Phase 1 models applied to the Phase 2 data in Table 3.12 reveals that:

1. The percent overall correct prediction was only slightly higher when the Phase 2 models were actually estimated than when applying the Phase 1 models to the Phase 2 data.

2. The percent correct prediction for severe and non-severe injuries, with the exception of Near NPCD and Far NPCD, were not significantly different when the Phase 2 data models were actually estimated and when directly applying the Phase 1 models to the Phase 2 data.

Figures 3.23 to 3.26 show the histograms of the probability, $\hat{p}_i$, of the occurrence of a non-severe injury as a function of Lateral Delta V and Age for Near PCD, Far PCD, Near NPCD and Far NPCD based on Equations 3-45 to 3-48. Each figure consists of two histograms, one for non-severe injuries and the other for severe injuries. Both histograms have the same axes, one represents the estimated probability of a non-severe injury and the other the number of cases with the particular $\hat{p}_i$ values. A $\hat{p}_i$ value greater than 0.5 would imply the occurrence of a non-severe injury while a value less than 0.5 would imply the occurrence of a severe injury. For Near PCD, Far PCD, Near NPCD and Far NPCD, the prediction of non-severe injuries was indeed very good, that is the models were predicting correctly almost all of the time. For severe injuries, however, the models were not doing quite as well, the Near PCD model was only predicting correctly on the average of about 45% of the time, the models for the other three subsets were correct about 20 to 30% of the time.

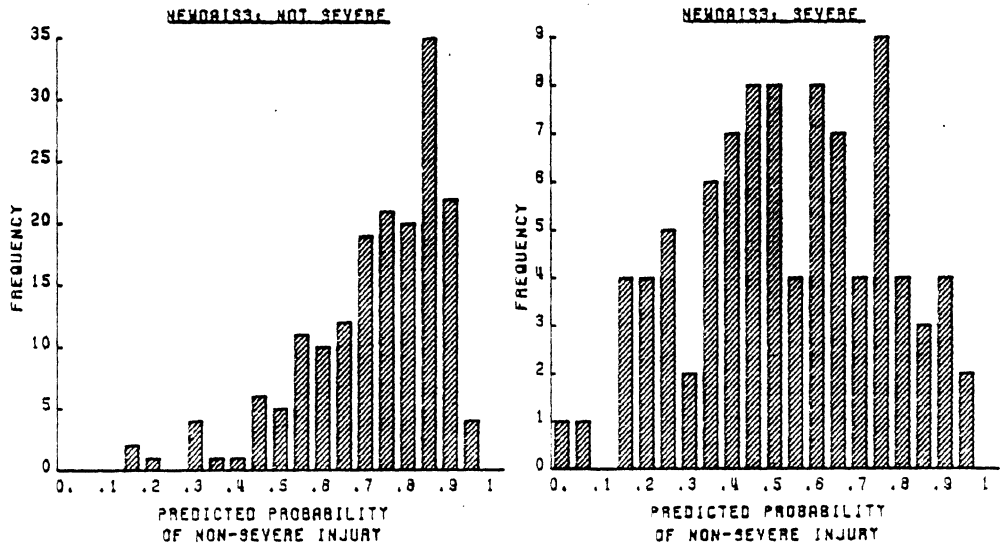

FIGURE 3.23 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For Near PCD Phase 2 Data - Side Impacts

The estimated logistic curves for each of these four subsets are shown in Figure 3.27. The logistic curves shows the estimated probability of an occupant receiving a severe injury ($1-\hat{p}_i$) as a function of Lateral Delta V alone; Age was fixed at 30 for these curves. It appears that of the two main subsets, Near PCD and Far PCD, the probability of a severe injury was higher in Near PCD for Lateral

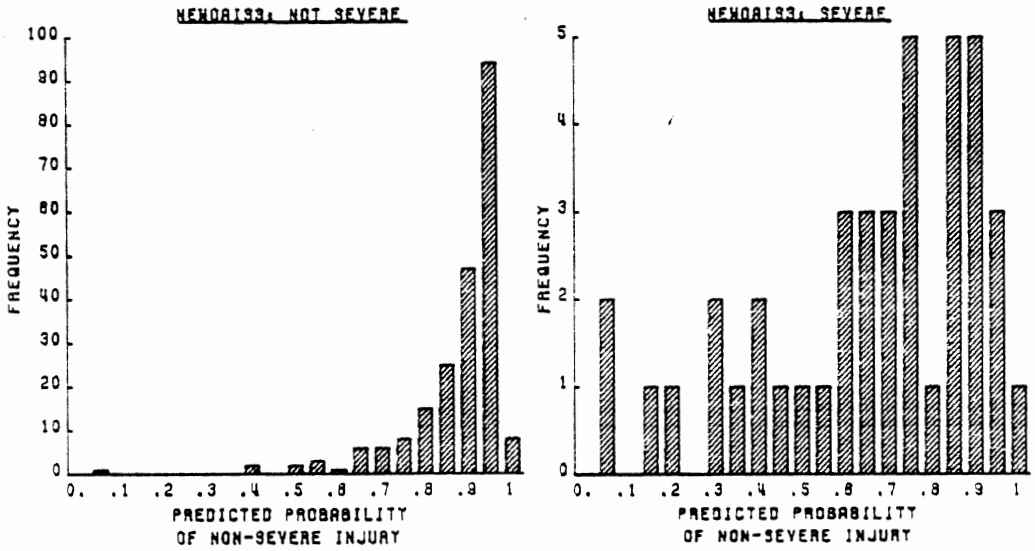

FIGURE 3.24 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For Far PCD Phase 2 Data - Side Impacts

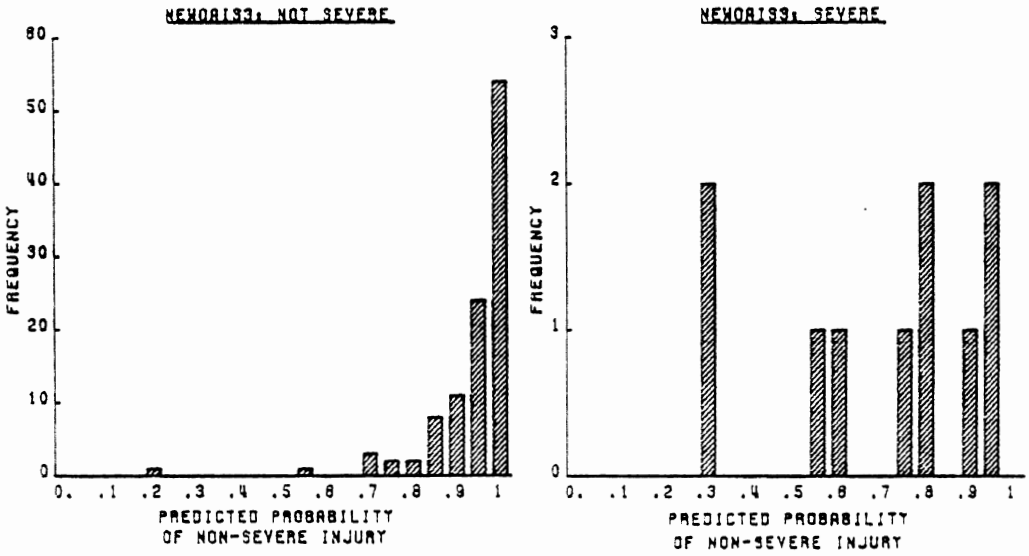

FIGURE 3.25 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For Near NPCD Phase 2 Data - Side Impacts

Delta V values smaller than 40 mph. As Lateral Delta V becomes larger than 40 mph the probability of a severe injury predicted by either model approaches one. The Near PCD model would predict a severe injury for Lateral Delta V greater than 20 mph and vice versa. The Far PCD model would predict a severe injury for Lateral Delta V greater than 28 mph and vice versa. Comparison of Figure 3.27 with Figure 3.11 (the Phase 1 logistic curves for the same subsets) indicated the similarities between the Near PCD models and the Far PCD models of the two phases. The

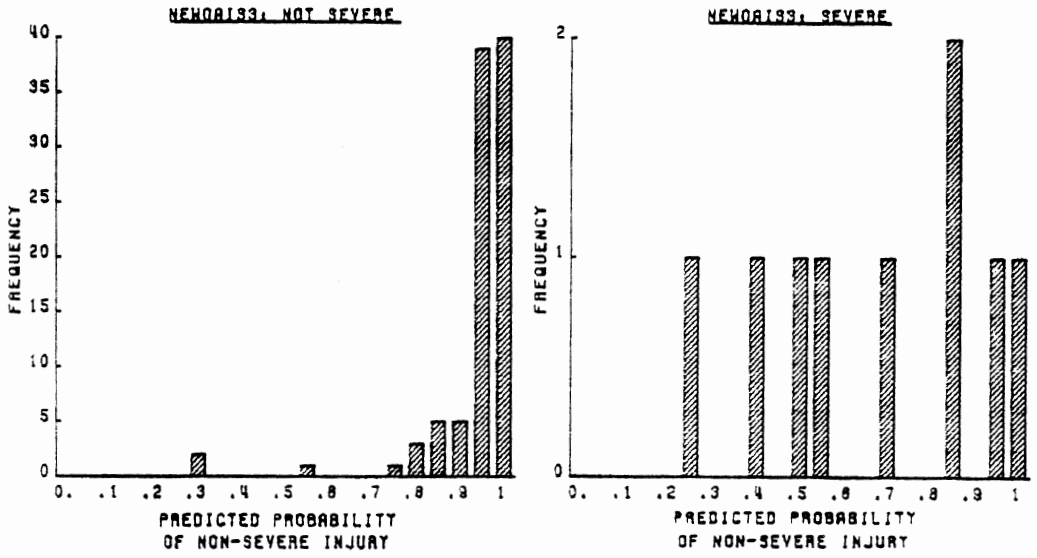

FIGURE 3.26 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For Far NPCD Phase 2 Data - Side Impacts

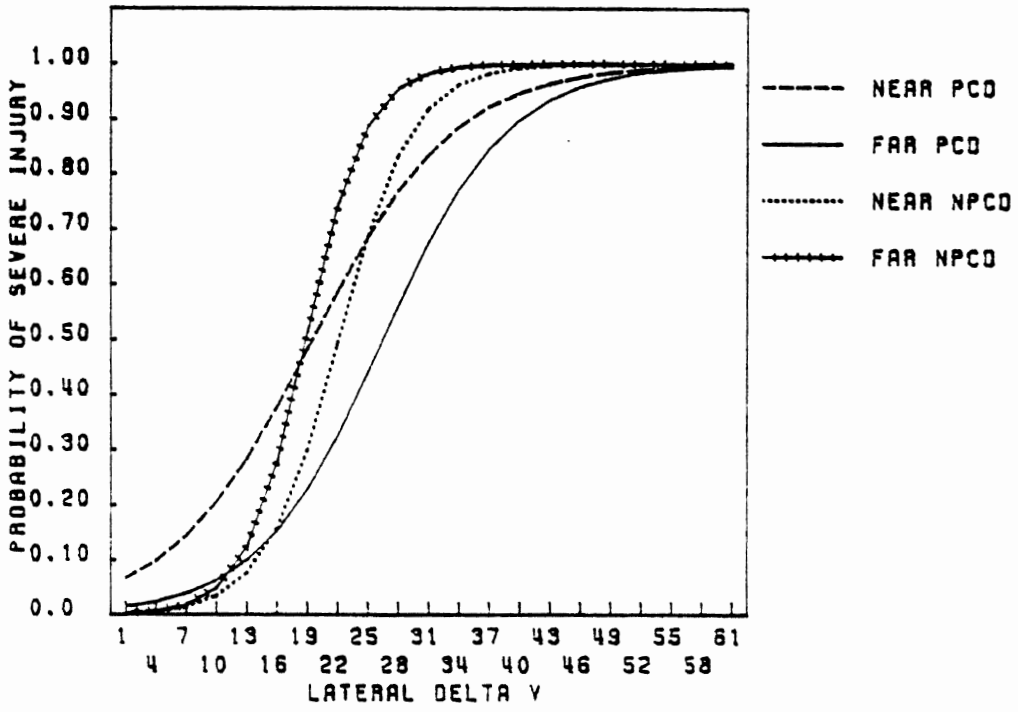

FIGURE 3.27 Logistic Curves of Two-Variable Model (Lateral Delta V, Age) For Side-Impact Subsets Phase 2 Data - Side Impacts

differences in the Near NPCD models of the two phases and the Far NPCD models of the two phases were considerable. These differences were not unexpected since the numbers of severe injuries in these subsets were

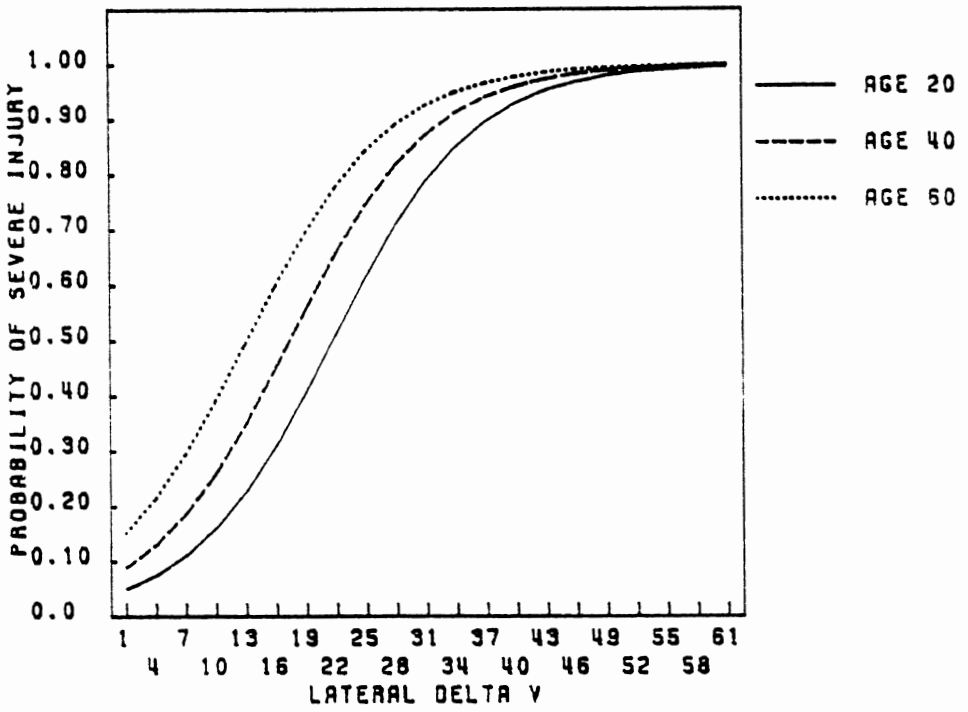

FIGURE 3.28 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Near PCD Phase 2 - Side Impacts

very small in both phases. The models of these subsets consequently had very wide confidence intervals. The effect of Age is illustrated for each subset in Figures 3.28 to 3.31. Each figure, representing a model for a particular subset, consists of three curves for Age of 20, 40 and 60. The curves are the plots of the probability of a severe injury ($1-\hat{p}_i$) as a function of Lateral Delta V. The effect of Age is comparable for the Near PCD, Near NPCD, and Far PCD, but negligible in the Far NPCD subset. The Age-Effect plots show, with the exception of Far NPCD, that given any value of Lateral Delta V an older occupant is expected to have a higher probability of a severe injury than a younger occupant. The confidence limits for each subset are illustrated in Figures 3.32 to 3.35. Each figure, representing a model for a particular subset, consists of three curves designating the upper bound, the lower bound and the estimated probability of a severe injury ($1-\hat{p}_i$). In general, the confidence limits are tighter for the PCD (passenger compartment damage) subsets. The narrower band of confidence limits indicates that the model has a smaller variance in predicting the probabilities of

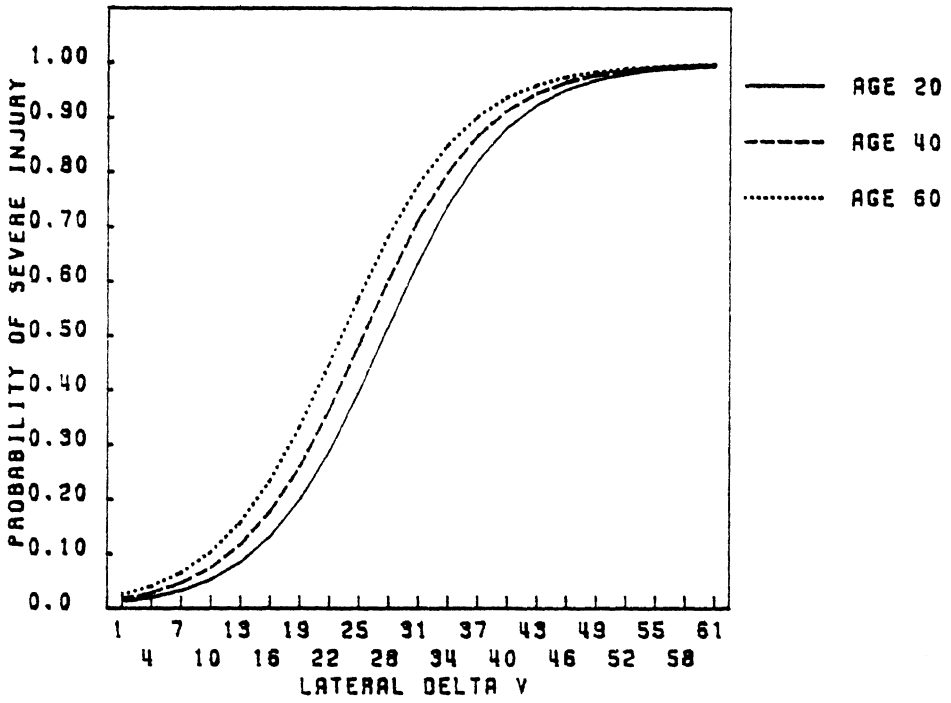

FIGURE 3.29 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Far PCD Phase 2 - Side Impacts

injury severity and therefore has a better chance of producing the similar results when different sets of data are analysed. Such repeatability is a desirable quality of a model. For the Near PCD and the Far PCD models, the largest variance in prediction occurs for the range of Lateral Delta V about 25 to 35 mph. For the models of the subsets with no passenger compartment damage the variances in prediction are very large, a result of a very small number of severe injuries in the samples (about 10% of total injuries). Their very large confidence limits are indicative of the models' inherent lack of stability. The Near and Far PCD subsets are compared in Figure 3.36. Substantial overlap in the confidence limits is seen as Lateral Delta V values become larger (greater than 25 mph). The difference between the two models in prediction is more considerable at low or moderate Lateral Delta V than at higher Lateral Delta V. The Near PCD model predicted a higher probability of a severe injury than the Far PCD model for a given value of Lateral Delta V.

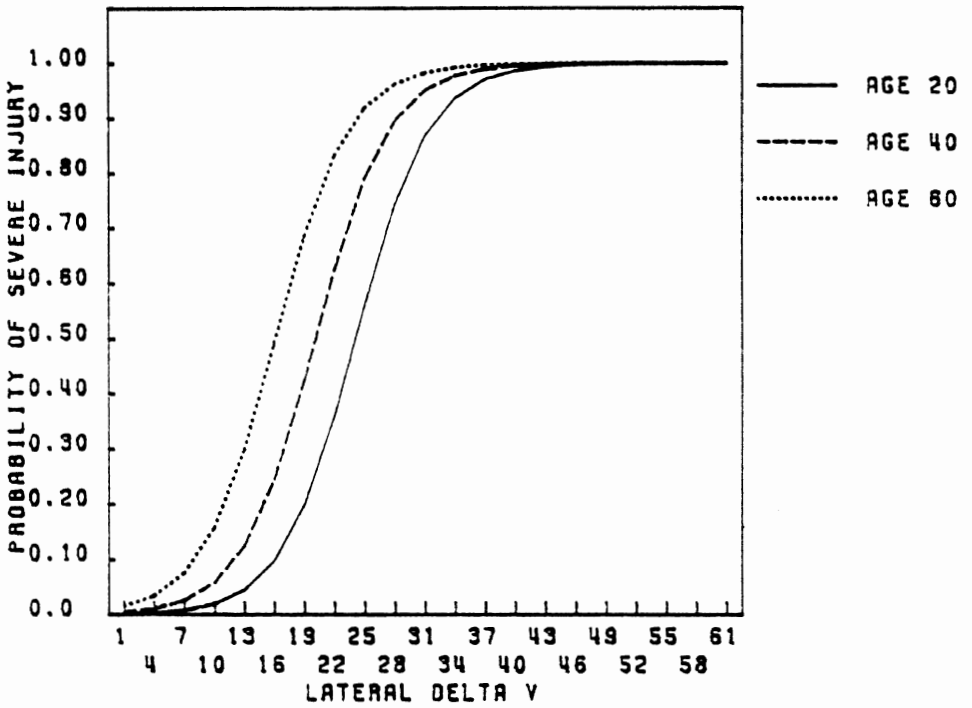

FIGURE 3.30 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Near NPCD Phase 2 - Side Impacts

Tables 3.16 to 3.17 summarizes the "outliers" by Body Region and Injury Type. Comparison of the "outliers" by Body Region in the Phase 1 and the Phase 2 data revealed that the body regions which tended to result in injury misprediction in both phases were comparable, such body regions were Chest, Abdomen, Forearm, and to a lesser extent Pelvic/Hip and the lower limbs. Injury Types which tended to result in injury misprediction were also comparable in both phases. These Injury Types were Rupture, Dislocation and Fracture. In both phases, fractures appeared to have occurred far more frequently than ruptures or dislocation. Table 3.18 summaries the outliers by Body Region for fractures only. Table 3.16 and Table 3.18 can be used together to give more information about the misprediction of fractures of specific body regions. For example, one can see that the majority of the mispredicted chest (34 out of 39) and pelvic-hip/thigh (12 out of 14) injuries in Table 3.16 were fractures. The analysis of the outliers in both Phase 1 and Phase 2 data indicated that the data could be pooled.

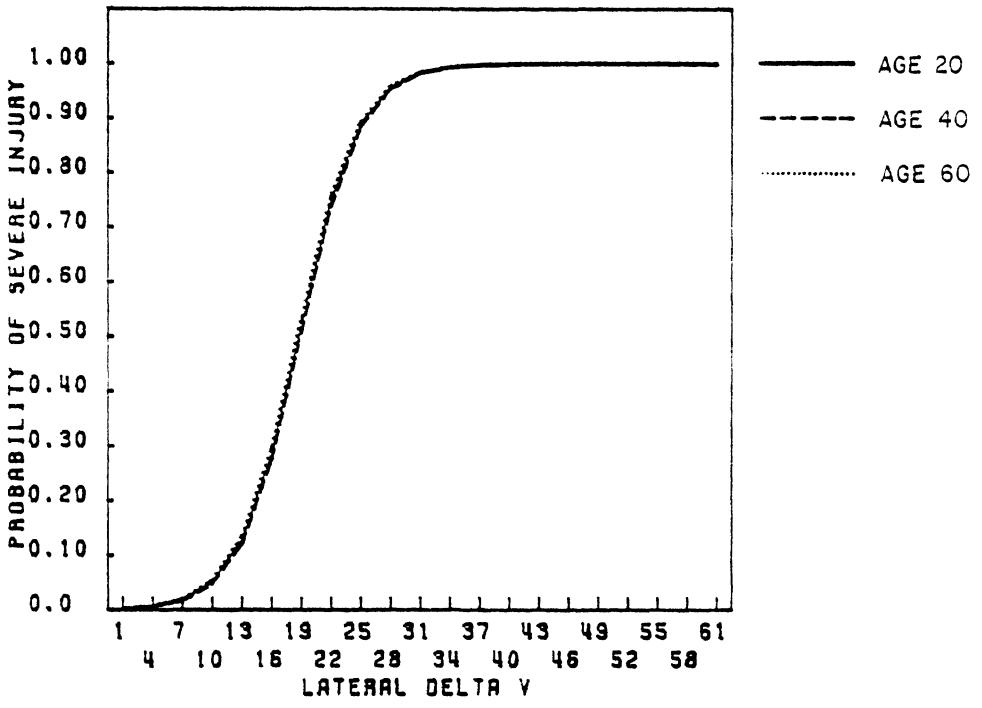

FIGURE 3.31 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Far NPCD Phase 2 - Side Impacts

The Phase 1 two-variable (Lateral Delta V and Age) models predicted injury severity almost as well when applied to the Phase 2 data. Furthermore, the Phase 1 models and the Phase 2 models did not yield appreciably dissimilar goodness of fit results. In general, the prediction of non-severe injuries in both phases was comparable. The prediction of severe injuries was also comparable for near-side occupants while that for far-side occupants showed somewhat more variability, which could be attributed to the relatively smaller sample size of the severe injuries of far-side occupants and/or to the fact that prediction of severe injuries had generally been, at least to date, tenuous. This seems to imply that the Phase 1 data and the Phase 2 data were similar so that they could be combined.

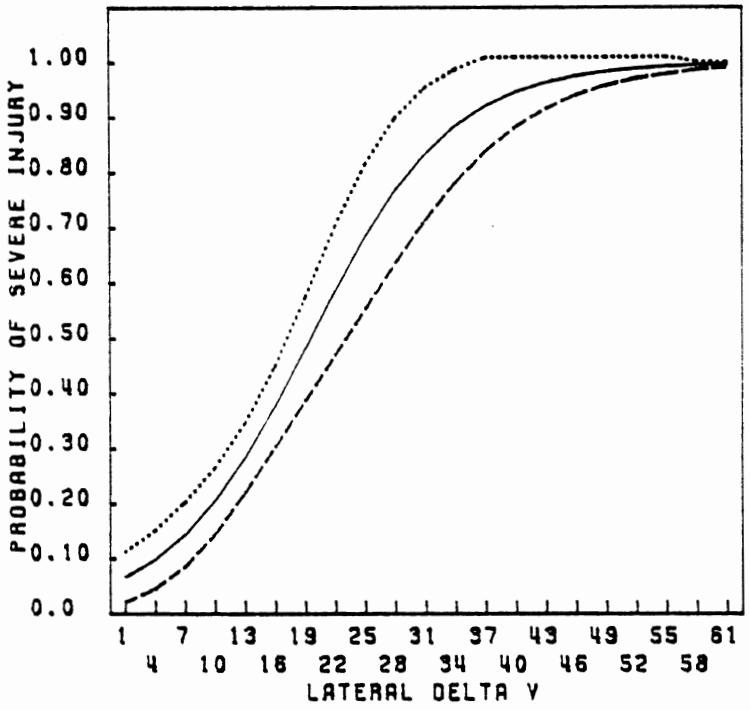

FIGURE 3.32 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Near PCD Phase 2 Data - Side Impacts

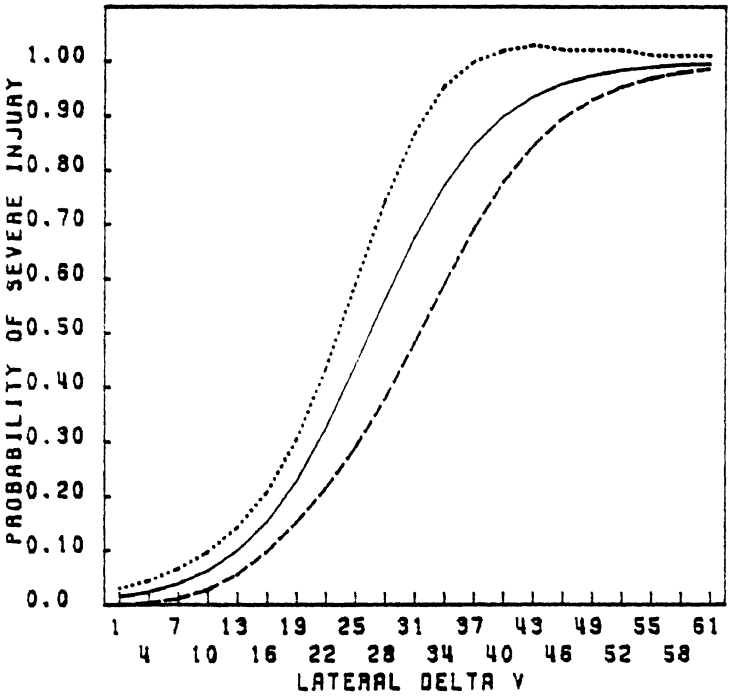

FIGURE 3.33 Confidence Interval of $\tilde{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Far PCD Phase 2 Data - Side Impacts

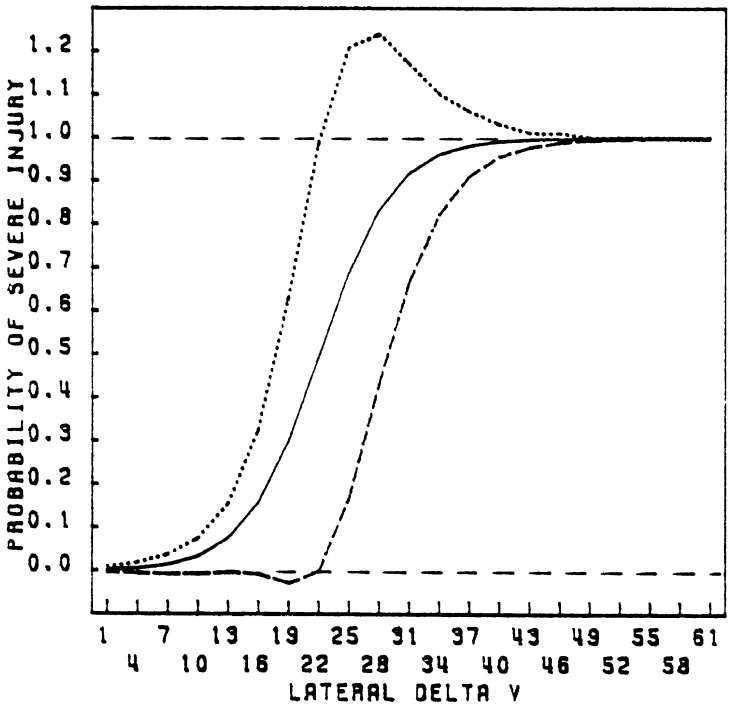

FIGURE 3.34 Confidence Interval of $\hat{p}_j$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Near NPCD Phase 2 Data - Side Impacts

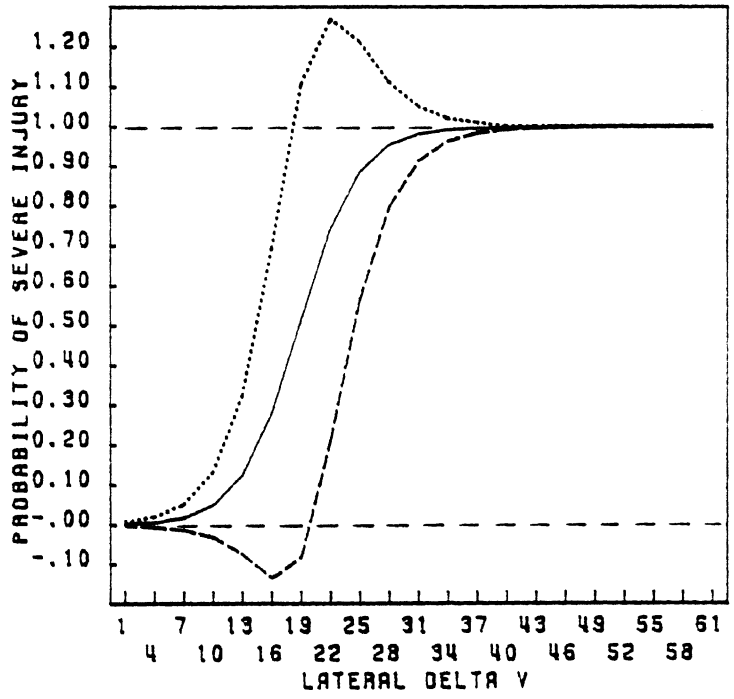

FIGURE 3.35 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Far NPCD Phase 2 Data - Side Impacts

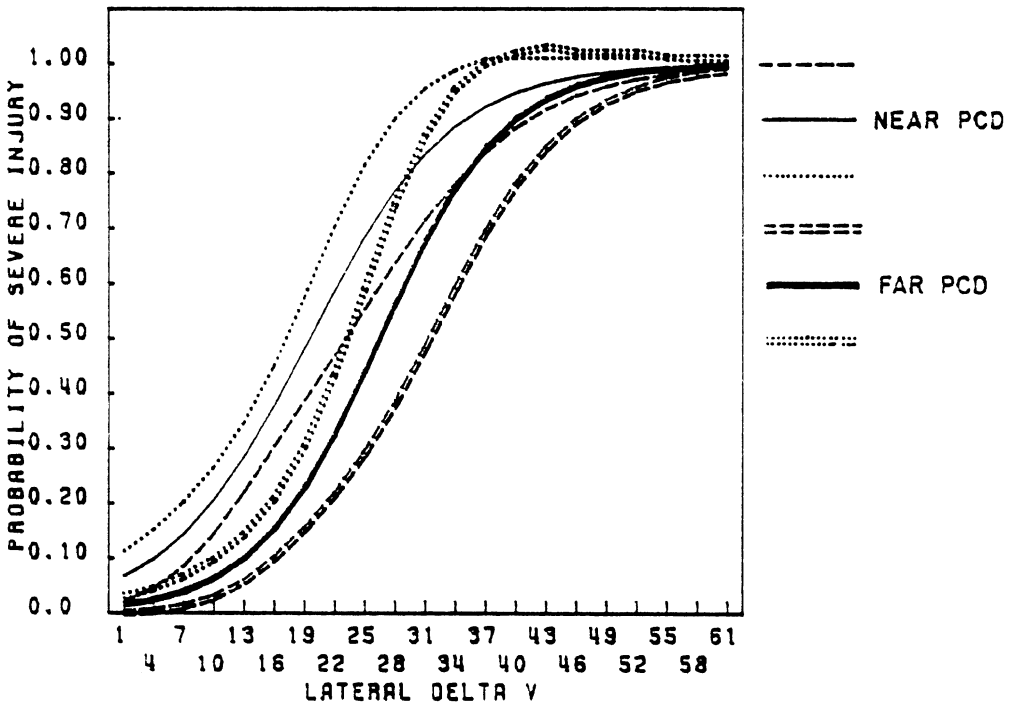

FIGURE 3.36 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Near PCD and Far PCD Phase 2 Data - Side Impacts

TABLE 3.16

List of Body Regions Associated with
Large Percent of Outliers

Phase 2 Data - Side Impacts

Subset	Body Regions* That Yield Large Percent of Misprediction	Total Number of Cases	Number of Misprediction
Near PCD	Forearm	6	5
	Abdomen	13	8
	Chest	38	17
	Pelvic/Hip,Thigh	27	11
Far PCD	Abdomen	8	5
	Chest	27	14
Near NPCD	Lower Leg/Ankle	2	2
	Chest	9	6
Far NPCD	Chest	3	2
All the Above Subsets Together	Forearms	9	6
	Abdomen	22	13
	Chest	76	39
	Pelvic/Hip,Thigh	38	14

*Body regions were ranked within subsets by the larger magnitude of misprediction proportions.

TABLE 3.17

List of Injury Types Associated with
Large Percent of Outliers

Phase 2 Data - Side Impacts

Subset	Injury Types Which Yield Large Percent of Misprediction	Total Number of Cases	Number of Mispredictions
Near PCD	Crushing	3	2
	Fracture	71	37
	Rupture	2	1
	Dislocation	2	1
Far PCD	Rupture	3	2
	Fracture	34	20
	Dislocation	2	1
Near NPCD	Fracture	11	8
Far NPCD	Abrasion	2	1
All the Above Subsets Together	Crushing	3	2
	Rupture	5	3
	Fracture	128	69
	Dislocation	5	2

*Injury Types were ranked within subsets by the higher percent of misprediction.

TABLE 3.18

Fractures Only
List of Body Regions Associated with
Large Percent of Outliers

Phase 2 Data - Side Impacts

Subset	Body Regions* Which Yield Large Percent of Mispredaiction	Total Number of Cases	Number of Misprediction
Near PCD	Neck	3	3
	Forearm	6	5
	Chest	25	14
	Thigh	6	3
	Face	2	1
	Pelvic/Hip	16	7
Far PCD	Chest	17	13
	Pelvic/Hip,Thigh	2	1
	Neck	2	1
	Face	2	1
Near NPCD	Lower Leg,Ankle	2	2
	Chest	6	5
Far NPCD	Chest	3	2
	Neck	2	1
All the Above Subsets Together	Forearm	8	6
	Neck	7	5
	Chest	51	34
	Face	5	3
	Pelvic/Hip,Thigh	26	12
	Ankle/Foot	2	1

*Body regions were ranked within subsets by the large percent of misprediction.

3.3.3 Combining Phase 1 and Phase 2 Data. The modelling results of the Phase 1 data and the Phase 2 data on the side collisions had indicated, in general, the Phase 1 two-variable models predicted the Phase 2 data nearly as well; they also predicted the Phase 2 data similarly to the estimated Phase 2 two-variable models. This was further confirmed by the similarity in the "outliers" based on the models of both phases in terms of injury types and body regions and by the similarity in the histograms of the $\hat{p}_i$ values of both phases for all subsets. Combining of the data from both phases was therefore further investigated statistically and the statistical results are shown in Table 3.19. Details of this aspect of combining the data can be found in Section 3.1. In brief, the null hypothesis, H_0, is that one model will adequately describe the Phase 1 and Phase 2 data. The alternative hypothesis, H_1, is that two independent models are required to describe the different phases. The statistical test used is the Likelihood Ratio Statistic which is discussed in more detail in Section 3.1.2. The results in Table 3.19 indicate that the Phase 1 and Phase 2 data can be combined for all the side-collision subsets.

TABLE 3.19

Statistical Results In
Combining Phase 1 and Phase 2 Data
Side Impacts

Subset	-2Log L_0 [a]	-2Log L_1 [b]	LRS [c]	df
Near PCD	679.04	676.24	2.80	3
Far PCD	378.90	375.84	3.06	3
Near NPCD	124.17	121.42	2.75	3
Far NPCD	97.13	92.86	4.27	3
ALL NEAR	849.45	847.52	1.93	3
ALL FAR	477.55	473.28	4.27	3

[a] L_0 is the likelihood of the data under the null hypothesis

[b] L_1 is the likelihood of the data under the alternative hypothesis

[c] LRS is asymptotically chi-square with df specified

Table 3.20 shows the number of cases of the combined Phase 1 and Phase 2 data with valid NEWOAIS3 codes, Lateral Delta V and Age.

TABLE 3.20

Descriptive Statistics for Key Variables
in the Side Impact Subsets

Phases 1 and 2 - Side Impacts

Subset	Sample Size	Proportions of Severe Injuries (%)	Lateral Delta V			Age		
			Range	Mean	S.D.	Range	Mean	S.D.
Near PCD	883	35.6	2-57	13.2	7.4	0-90	31	18.8
Far PCD	870	17.4	1-52	13.6	7.8	0-98	31	18.5
Near NPCD	419	6.2	1-36	8.3	4.4	0-85	33	18.5
Far NPCD	416	5.8	1-28	8.6	4.4	0-85	32	18.6

Figure 3.37 shows the cumulative distribution of Lateral Delta V for all subsets. The cumulative curves of Lateral Delta V for Near PCD and Far PCD are almost identical. They are different from those for Near NPCD and Far NPCD, whose range of Lateral Delta V was considerably smaller. The results of the model estimation for the combined Phase 1 and Phase 2 data are shown in Equations 3-51 to 3-56.

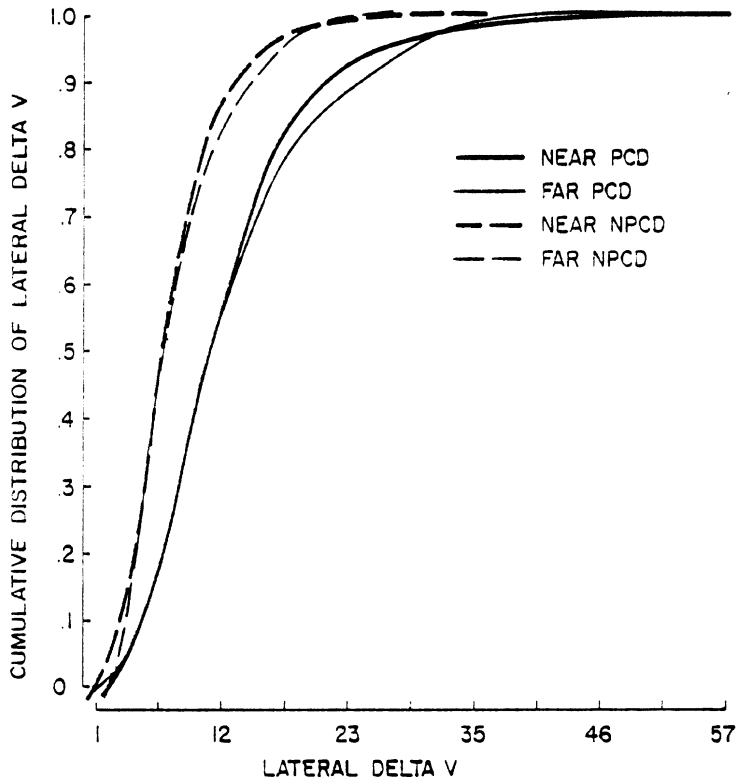

FIGURE 3.37 Cumulative Distributions of Lateral Delta V For Side-Impact Subsets Phases 1 and 2 - Side Impacts

Estimated Models with Lateral Delta V and Age

Near PCD (N=650, LRS=147.51, DF=2)

(3-51) $\hat{p}_i = F(2.1426 - 0.0926X_1 - 0.0152X_2)$

Far PCD (N=639, LRS=154.86, DF=2)

(3-52) $\hat{p}_i = F(3.1356 - 0.1079X_1 - 0.0124X_2)$

Near NPCD (N=329, LRS=32.03, DF=2)

(3-53) $\hat{p}_i = F(3.5282 - 0.1014X_1 - 0.0248X_2)$

Far NPCD (N=325, LRS=30.43, DF=2)

(3-54) $\hat{p}_i = F(3.6222 - 0.1401X_1 - 0.0111X_2)$

Near All (N=979, LRS=234.24, DF=2)

(3-55) $\hat{p}_i = F(2.5262 - 0.1074X_1 - 0.0158X_2)$

Far All (N=964, LRS=206.93, DF=2)

where

$\hat{p}_i$ is the probability of a non-severe injury,
F is the logistic distribution,
X_1 is Lateral Delta V,
X_2 is Age, and
LRS is the Likelihood Ratio Statistic.

The goodness of fit results of the estimated models for the combined data are shown in Table 3.21.

Figures 3.38 to 3.43 show the histograms of the $\hat{p}_i$ values of the six subsets based on Equations 3-51 to 3-56. Each figure, representing each model for a particular subset, consists of a pair of histograms, one for non-severe injuries and the other for severe injuries. The axes of both histograms are identical, one representing the estimated probability of a non-severe injury ($\hat{p}_i$) at a 0.05 interval and the other the number of cases with particular $\hat{p}_i$ values. The Phase 1 and Phase 2 combined two-variable (Lateral Delta V and Age) models indicated that, in general, the models for far-side occupants tended to predict overall

TABLE 3.21

Goodness of Fit

Severity = F(Lateral Delta V, Age)

Phases 1 and 2 - Side Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
Near PCD	434	216	75.2	91.9	41.7
Far PCD	545	94	89.1	97.6	39.4
Near NPCD	308	21	93.3	99.4	4.8
Far NPCD	309	16	94.8	99.4	6.3
ALL FAR	742	237	80.7	95.7	33.8
ALL NEAR	854	110	91.0	98.2	34.5

injury severity better than those for near-side occupants. Of the four independent subsets, Near PCD, Far PCD, Near NPCD and Far NPCD, the first subset had the lowest overall percent correct prediction (75% compared to 90% or over for the other three subsets). This is due to the near-perfect prediction of non-severe cases in the other three subsets. As with the Phases 1 and Phases 2 models, the prediction of non-severe injuries is very good for all subsets. The prediction of severe injuries, however, remains somewhat unsatisfactory.

The estimated logistic curves for the four subsets Near PCD, Far PCD, Near NPCD, and Far NPCD are shown in Figure 3.44. These curves show the effect of Lateral Delta V on the predicted probability of a severe injury ($1-\hat{p}_i$) when Age is held fixed at 30. The Near PCD curve looks quite different from the other three curves, particularly at the lower range of Lateral Delta V values. The curves suggest that occupants of the Near PCD subset had considerably higher estimated probabilities of severe injures than those in the other three subsets. The logistic curves of the other three subsets - Far PCD, Far NPCD, and Near NPCD look very similar to one another. The effect of Age is shown in Figures 3.45 to 3.48. All four subsets now show comparable effects

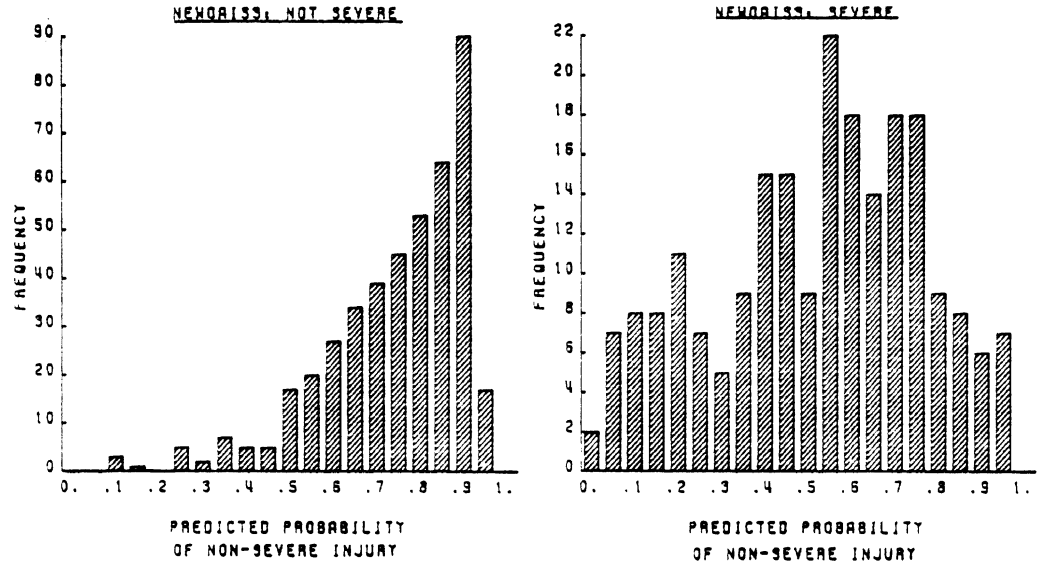

FIGURE 3.38 Histograms of Two-Variable Model (Lateral Delta V, Age) For Near PCD Phases 1 and 2 - Side Impacts

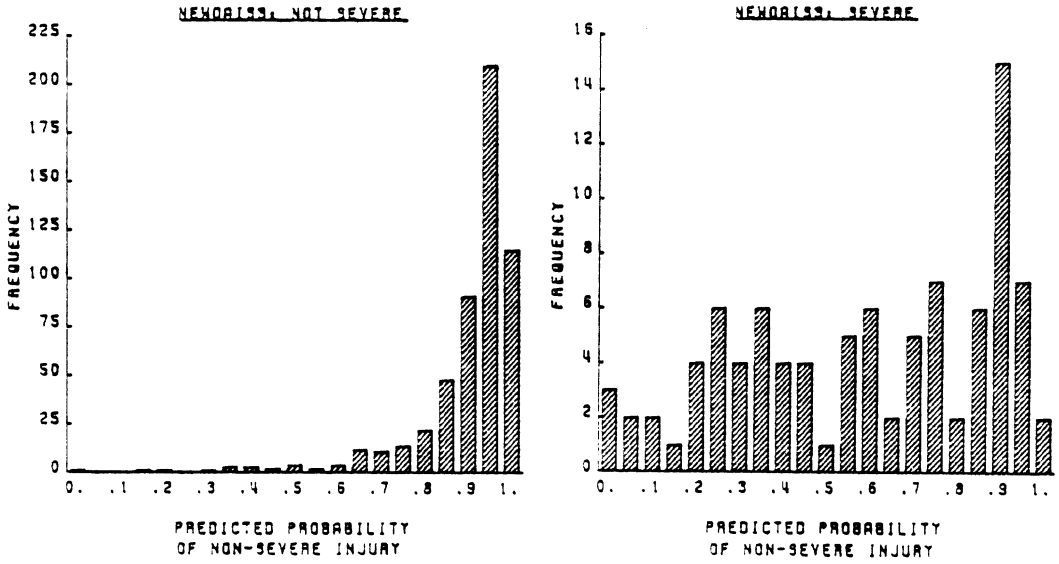

FIGURE 3.39 Histograms of Two-Variable Model (Lateral Delta V, Age) For Far PCD Phases 1 and 2 - Side Impacts

of Age. Each of the Age-Effect figures consists of three curves for Age at 20, 40 and 60. The age effect tends to be move pronounced for the Far PCD and the Near PCD models than for the models involving no passenger compartment damage. Furthermore, in the former two subsets, the age effect is most prominent at moderate Lateral Delta V values (about 10 to 30 mph). For all subsets, older occupants are expected to show less resistance to severe injuries than the younger counterparts. Confidence limits as a function of Delta V are shown in Figures 3.49 -

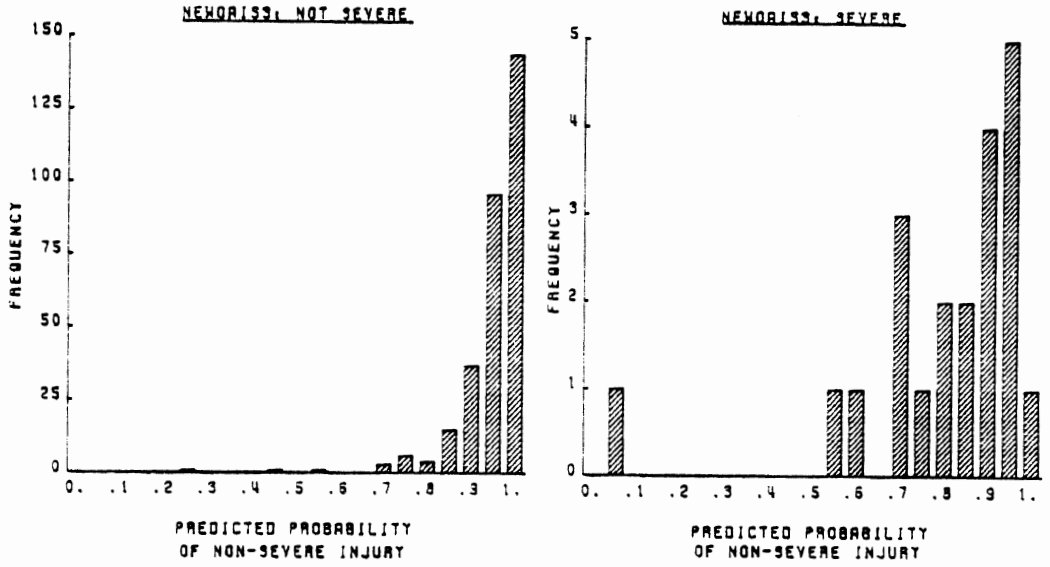

FIGURE 3.40 Histograms of Two-Variable Model (Lateral Delta V, Age) For Near NPCD Phases 1 and 2 - Side Impacts

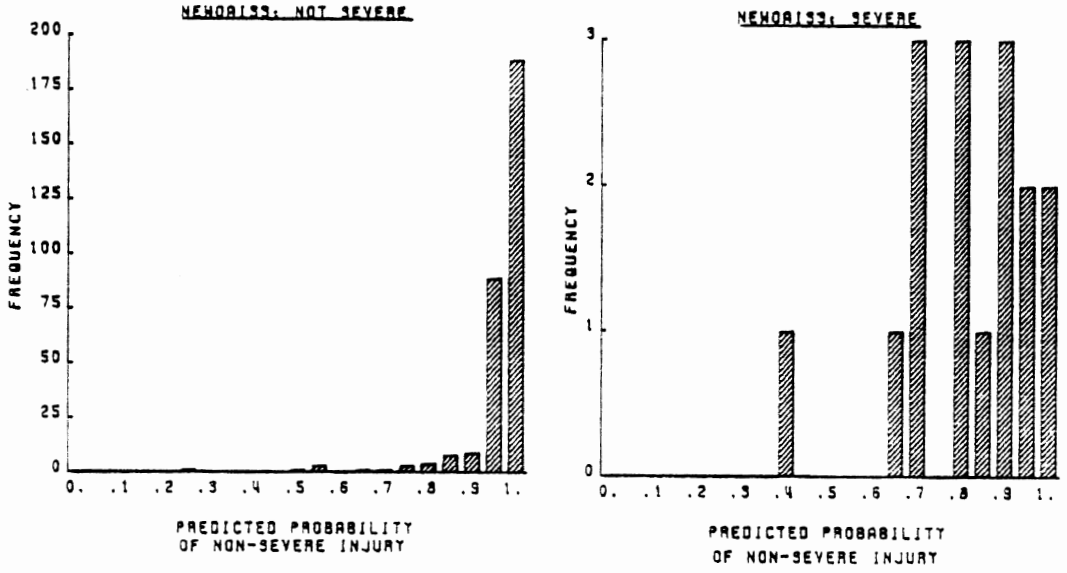

FIGURE 3.41 Histograms of Two-Variable Model (Lateral Delta V, Age) For Far NPCD Phases 1 and 2 - Side Impacts

3.52. Each figure, representing a model for a particular subset, consists of three curves which are the upper bound, the lower bound and the estimated probability of a severe injury $(1-\hat{p}_i)$. For the Near PCD model, the confidence interval approaches zero as Lateral Delta V approaches a value of about 50 mph. For Far PCD, Near NPCD, and Far NPCD the confidence intervals approach zero as Lateral Delta V becomes very small or very large. The confidence limits of the Near PCD and the Far PCD models are considerably smaller than those of the Near NPCD and

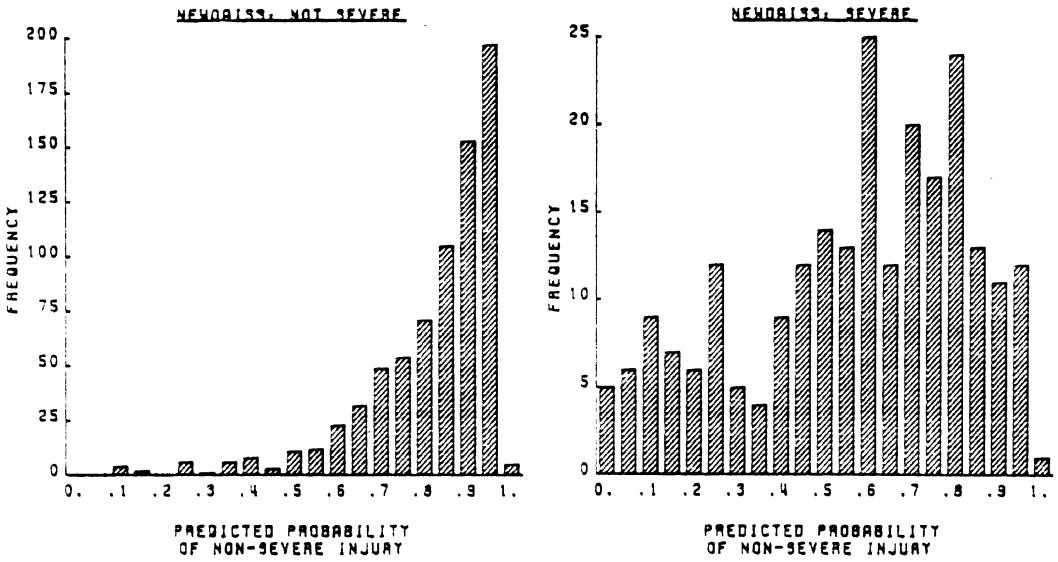

FIGURE 3.42 Histograms of Two-Variable Model (Lateral Delta V, Age) For All Near Phases 1 and 2 - Side Impacts

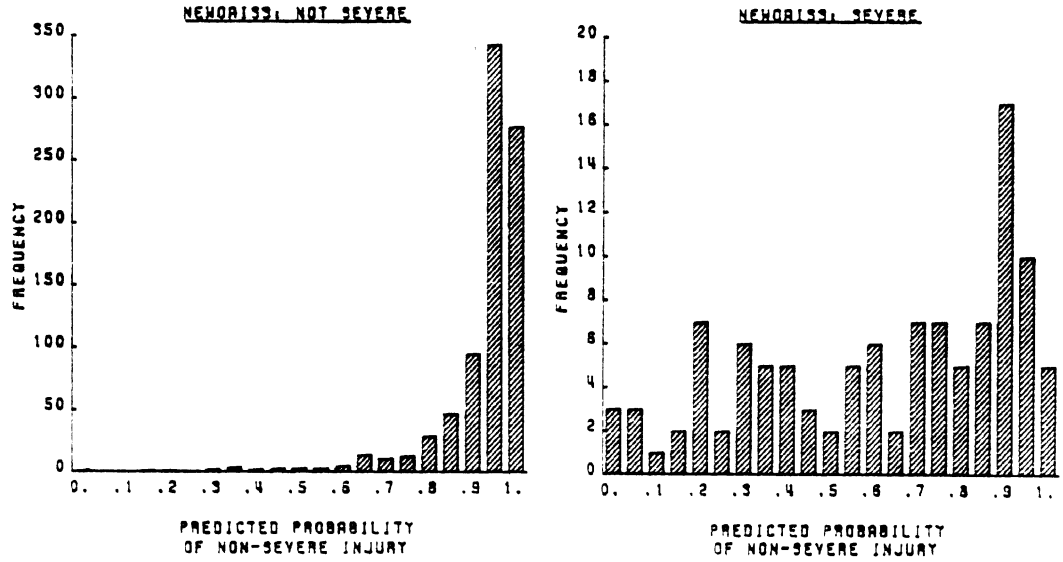

FIGURE 3.43 Histograms of Two-Variable Model (Lateral Delta V, Age) For All Far Phases 1 and 2 - Side Impacts

the Far NPCD models. This is due to the much smaller number of severe injuries in Near NPCD and Far NPCD (less than 10% of total injuries). Figure 3.53 compares the Near and Far PCD subsets. The figure shows that the difference between the Near PCD model and the Far PCD models is the prediction of severity is more considerable when Lateral Delta V values are low or moderate (up to 30mph). A small overlap of the two confidence intervals occurs at Lateral Delta V of greater than 25 mph, the range at which the estimated probability of a severe injury is

approaching a value of one. Figure 3.53 indicates that the Near PCD and the Far PCD models are dissimilar.

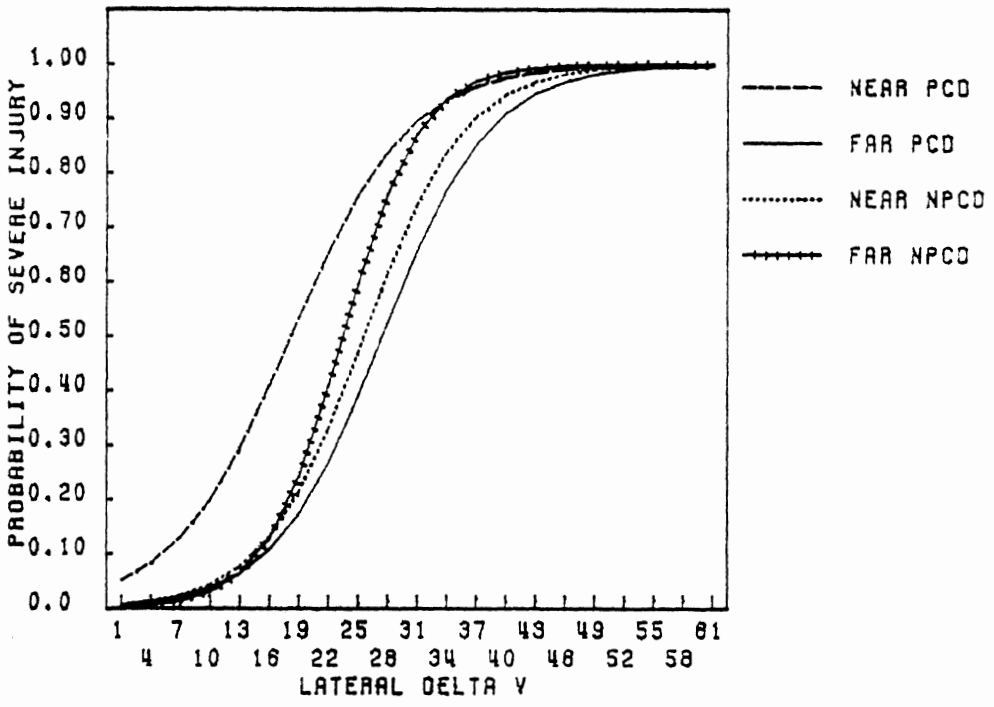

FIGURE 3.44 Logistic Curves of Two-Variable Models (Lateral Delta V, Age) For Side-Impact Subsets Phases 1 and 2 - Side Impacts

The modelling results so far from the combined data suggested that the non-passenger compartment damage subsets (i.e., Near NPCD and Far NPCD) were not very different. Furthermore, these two subsets appeared to be far more similar to Far PCD than to Near PCD both in terms of model estimation and goodness of fit results. In fact, further statistical investigation of the two-variable models of these four independent subsets revealed that the combined Phase 1 and Phase 2 data can be pooled across the subsets as shown by the statistical results in Tables 3.22 and 3.23. Collapsing of the individual subsets in this manner was indeed desirable since it would increase the sample size of severe injuries and therefore make the subsequent models more stable. In addition, by combining Far PCD with Far NPCD and Near NPCD to form one subset while retaining Near PCD as another subset, the ranges of the key independent variable, especially Lateral Delta V, became more

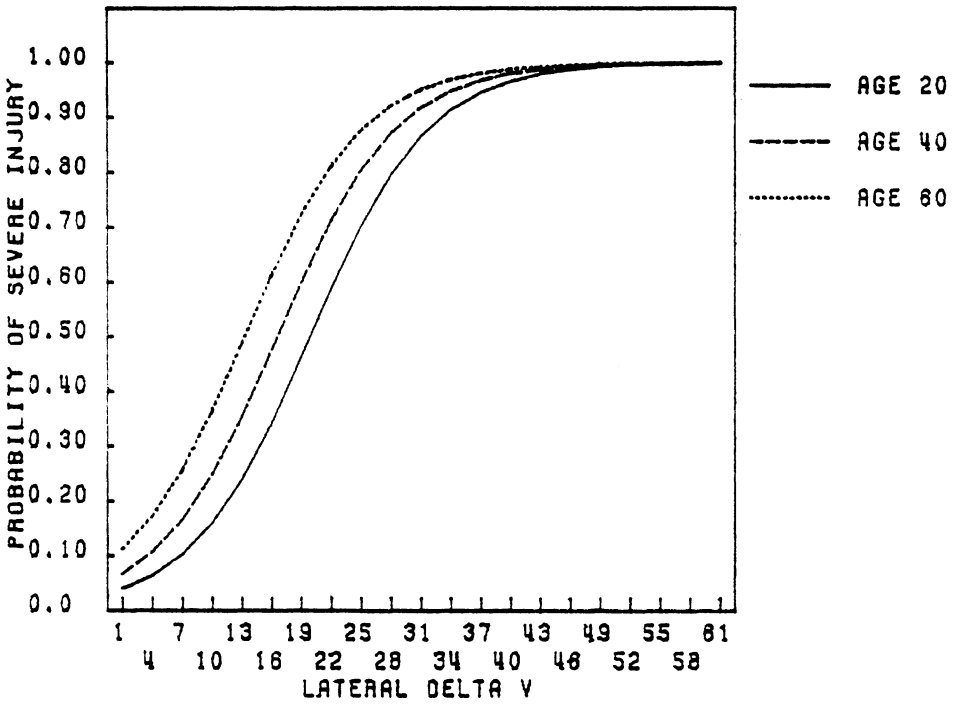

FIGURE 3.45 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Near PCD Phases 1 and 2 - Side Impacts

comparable across these two new subsets. Comparison of the models for different subsets now would become more meaningful.

Therefore the final subsets used in the combined Phase 1 and Phase 2 were:

1. Near PCD which refers to near-side occupants with passenger compartment damage.
2. Far Occ + Near NPCD which includes all far side occupants and near-side occupants with non-passenger compartment damage.

The modelling results of the new subsets which have Lateral Delta V and Age as the independent variables are shown below.

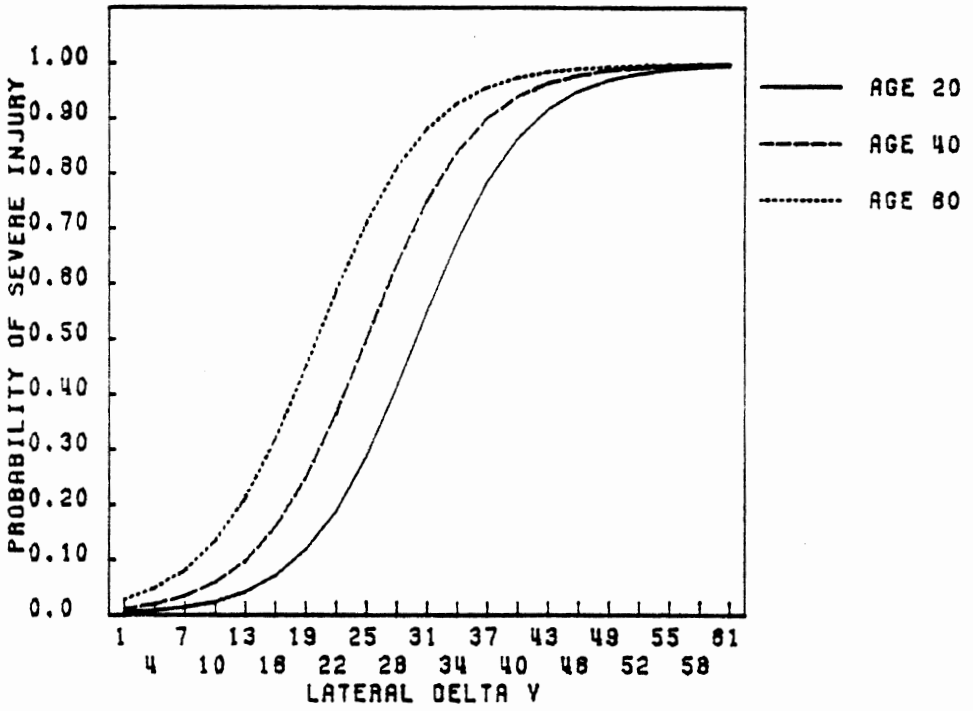

FIGURE 3.46 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Far PCD Phases 1 and 2 - Side Impacts

Estimated Models of the New Subsets with Lateral Delta V and Age

Near PCD (N=650, LRS=147.51 , DF=2)

(3-57) $\hat{p}_i = F(2.1426 - 0.0926X_1 - 0.0152X_2)$

Far OCC. + Near NPCD (N=1293, LRS=242.14, DF=2)

(3-58) $\hat{p}_i = F(3.2672 - 0.1104X_1 - 0.0146X_2)$

where

$\hat{p}_i$ is the estimated probability of a non-severe injury,
F is the logistic distribution,
X_1 is Lateral Delta V,
X_2 is Age, and
LRS is the Likelihood Ratio Statistic.

The goodness of fit results of Equations 3-57 and 3-58 are shown in Table 3.24 and Figures 3.54 and 3.55 show the histograms of the $\hat{p}_i$ values for the Near PCD model and the (Far Occ + Near NPCD models)

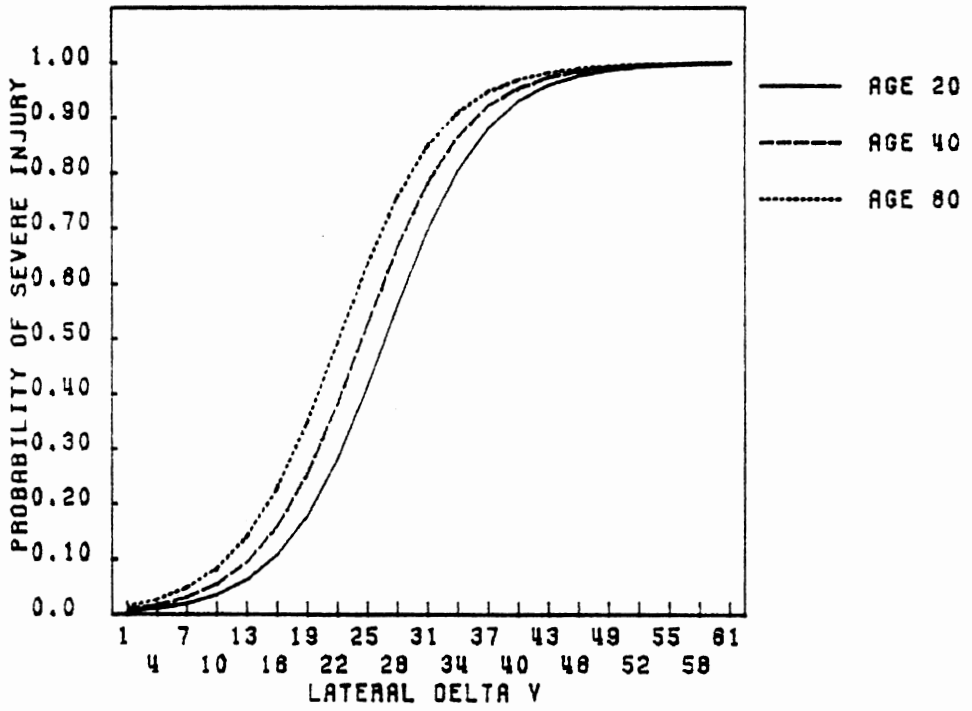

FIGURE 3.47 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Near NPCD Phases 1 and 2 - Side Impacts

respectively. Again, the models predicted non-severe injuries almost perfectly but the prediction of severe injuries was somewhat unsatisfactory.

The estimated logit curves for the combined subsets are shown in Figure 3.56. Confidence intervals are shown on this plot. The confidence intervals of both models are quite tight, particularly at either very low or very high values of Lateral Delta V. The difference in the Near PCD model and the (Far Occ + Near NPCD) model in predicting the severity is that the former will, given a value of Lateral Delta V, result in a higher probability of a severe injury almost all of the time. This is especially true for Lateral Delta V less than about 30 mph. Comparison of Figure 3.57 and 3.58 illustrates that the effect of Age is very similar in the two final subsets. Each figure, consisting of three curves for Age of 20, 40 and 60, indicates that older occupants, in general, show higher probabilities of severe injuries than younger occupants. Confidence limits are shown separately for each subset in Figures 3.59 and 3.60.

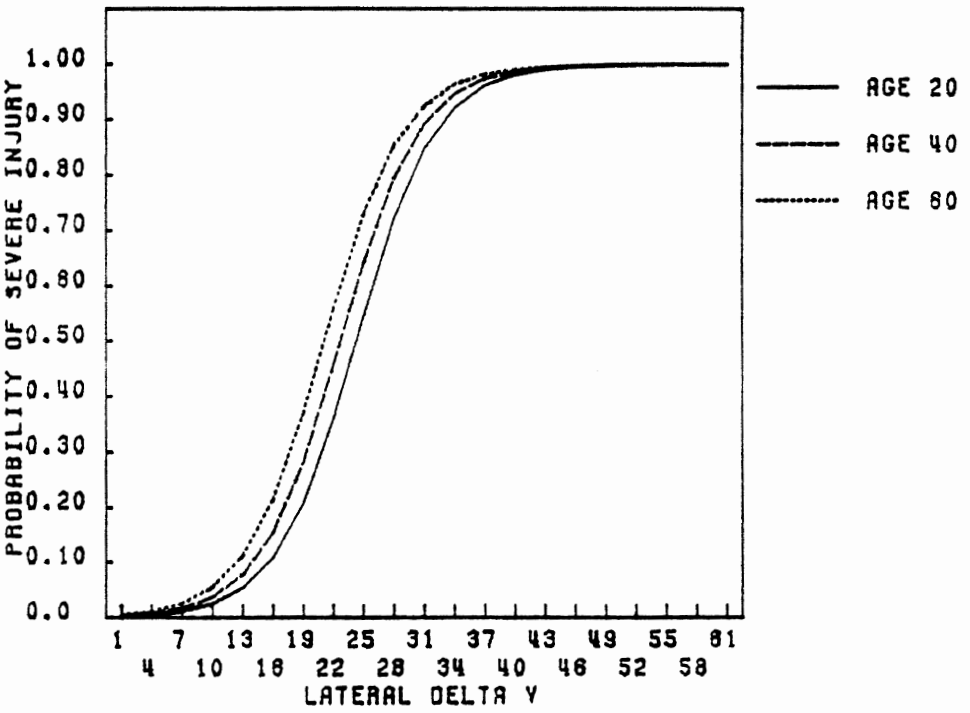

FIGURE 3.48 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Far NPCD Phases 1 and 2 - Side Impacts

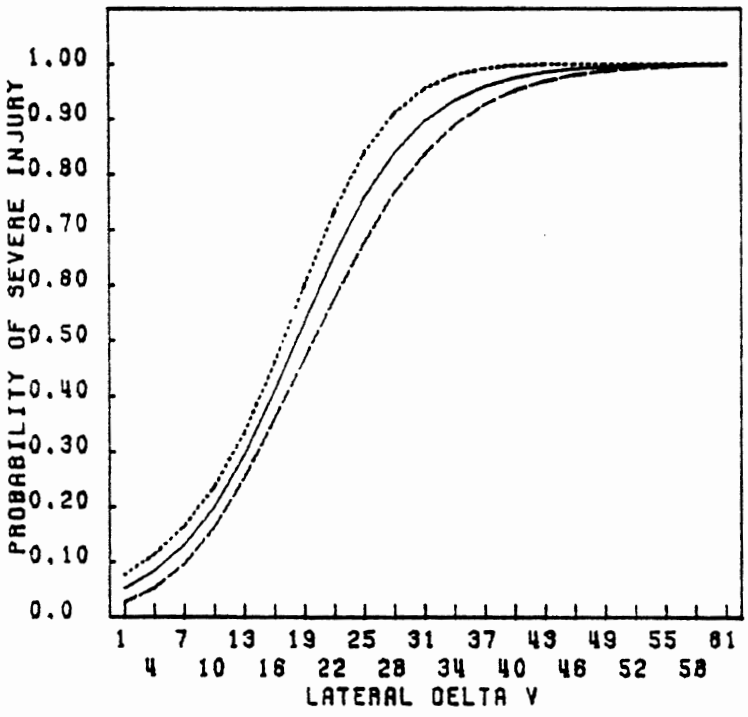

FIGURE 3.49 Confidence Interval of $\hat{p}_i$ of Two-Variable Models (Lateral Delta V, Age) at Age 30 For Near PCD Phase 1 and 2 - Side Impacts

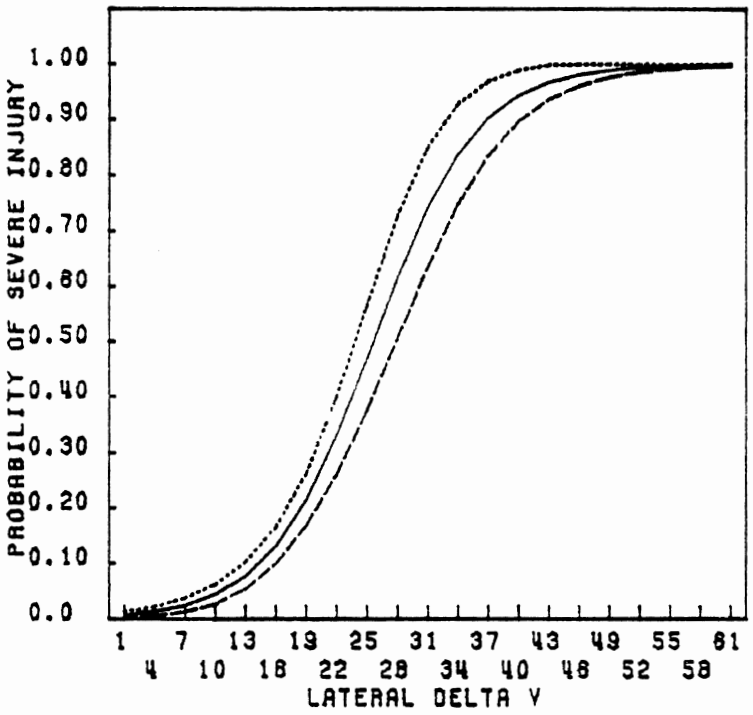

FIGURE 3.50 Confidence Interval of $\hat{p}_i$ of Two-Variable Models (Lateral Delta V, Age) at Age 30 For Far PCD Phase 1 and 2 - Side Impacts

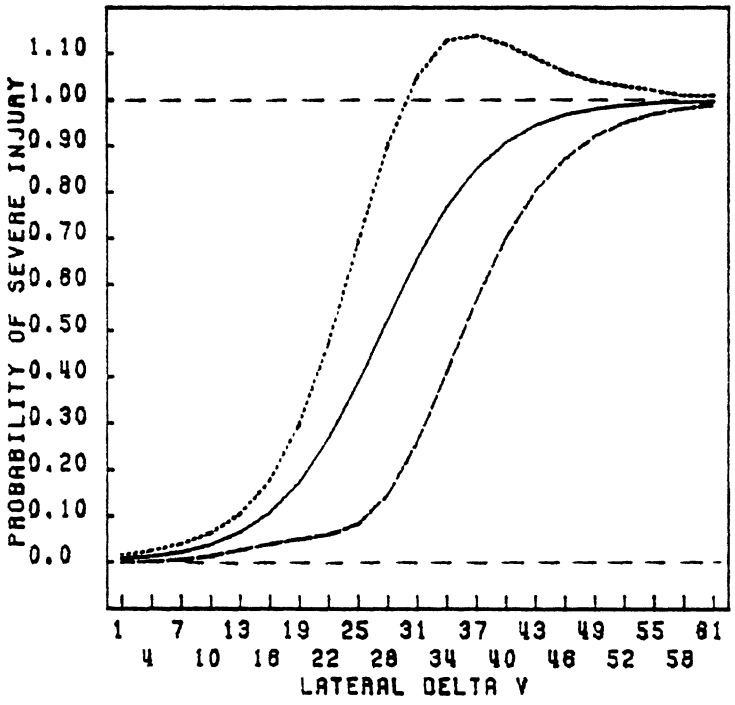

FIGURE 3.51 Confidence Interval of $\hat{p}_i$ of Two-Variable Models (Lateral Delta V, Age) at Age 30 For Near NPCD Phase 1 and 2 - Side Impacts

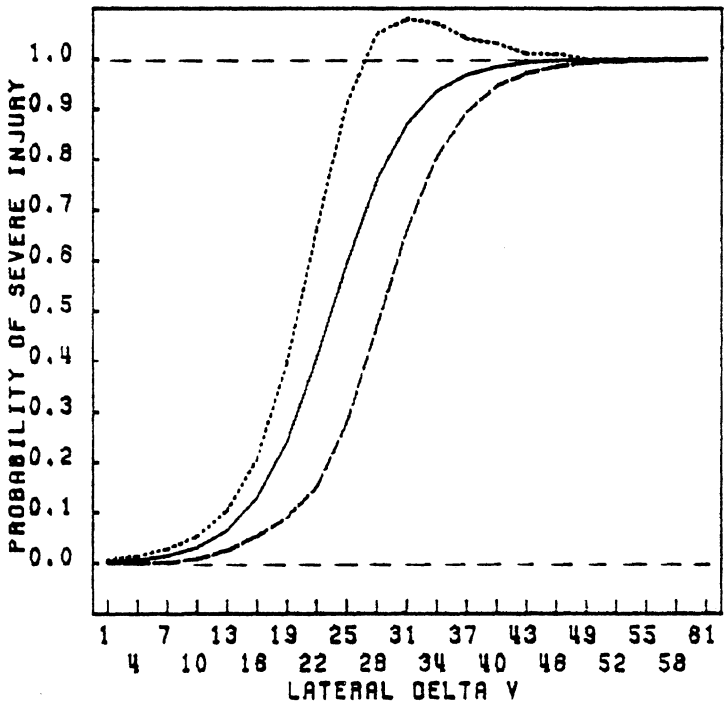

FIGURE 3.52 Confidence Interval of $\hat{p}_i$ of Two-Variable Models (Lateral Delta V, Age) at Age 30 For Far NPCD Phase 1 and 2 - Side Impacts

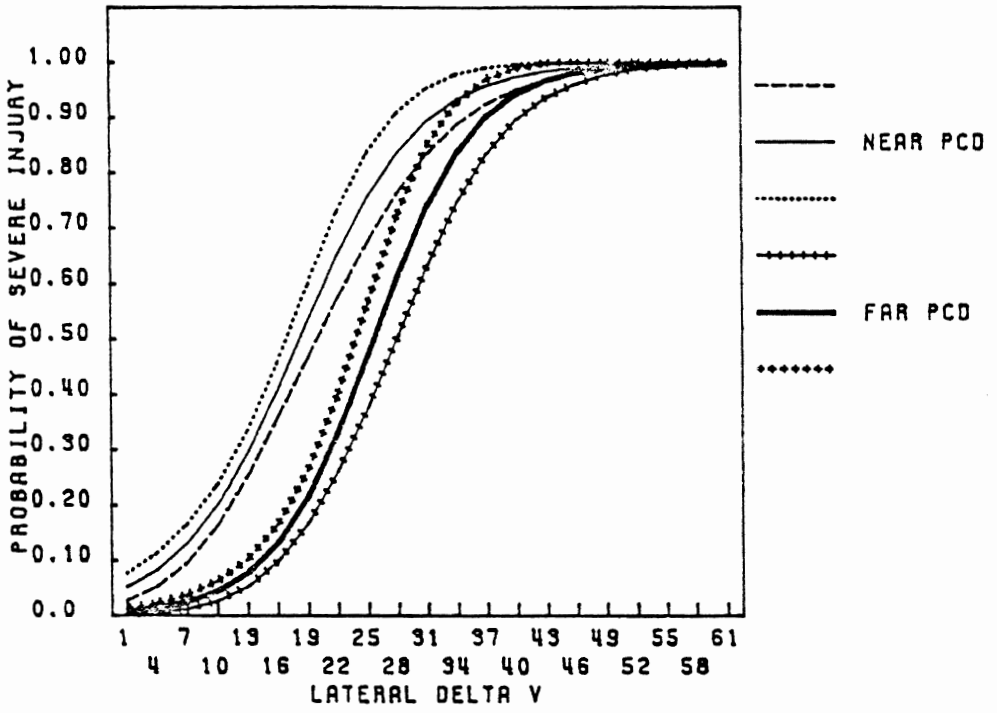

FIGURE 3.53 Confidence Interval of $\hat{p}_i$ of Two-Variable Models (Lateral Delta V, Age) at Age 30 For Near PCD and Far PCD Phase 1 and 2 - Side Impacts

TABLE 3.22

Statistical Results
Combining Far PCD, Far NPCD and Near NPCD

Phases 1 and 2 Side Impacts

Hypothesis	-2Log L	df.
H_0	605.97	3
H_1	600.20	9
	Chi-Square = 5.77	6

where H_0: 3 subsets have the same model

H_1: 3 subsets have different models

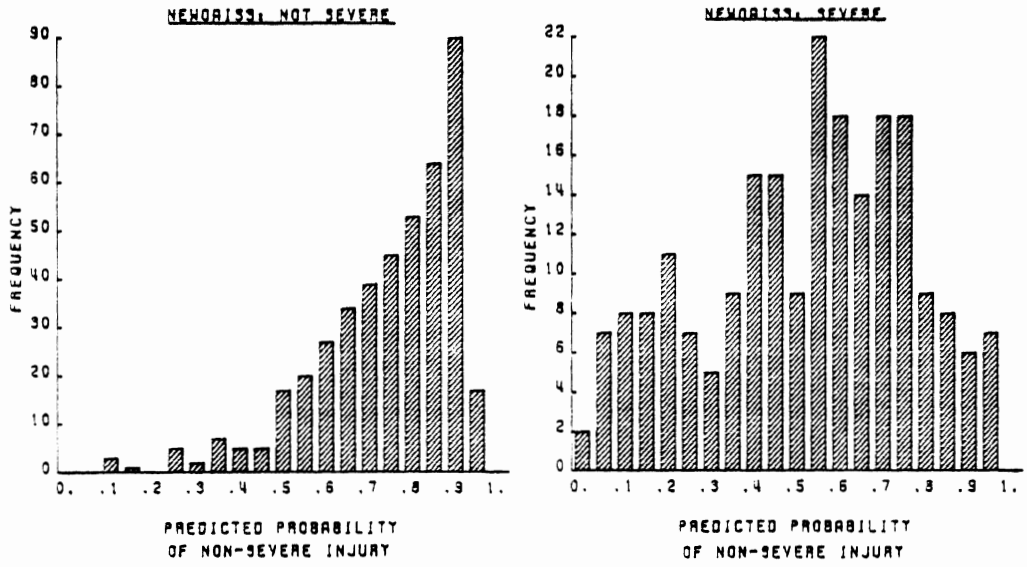

FIGURE 3.54 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For Near PCD Phases 1 and 2 - Side Impacts

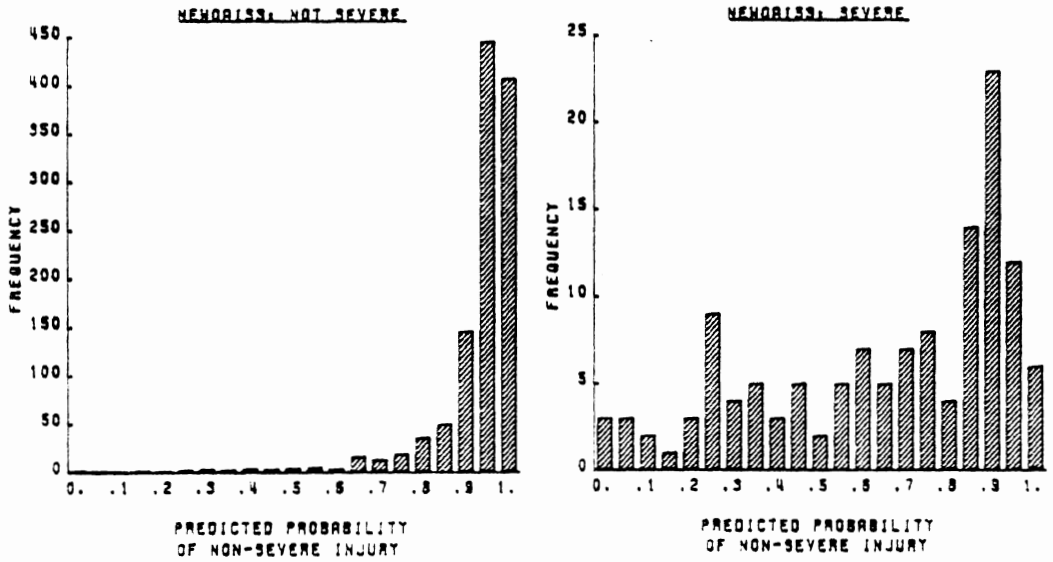

FIGURE 3.55 Histograms of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) For (Far Occ + Near NPCD) Phases 1 and 2 - Side Impacts

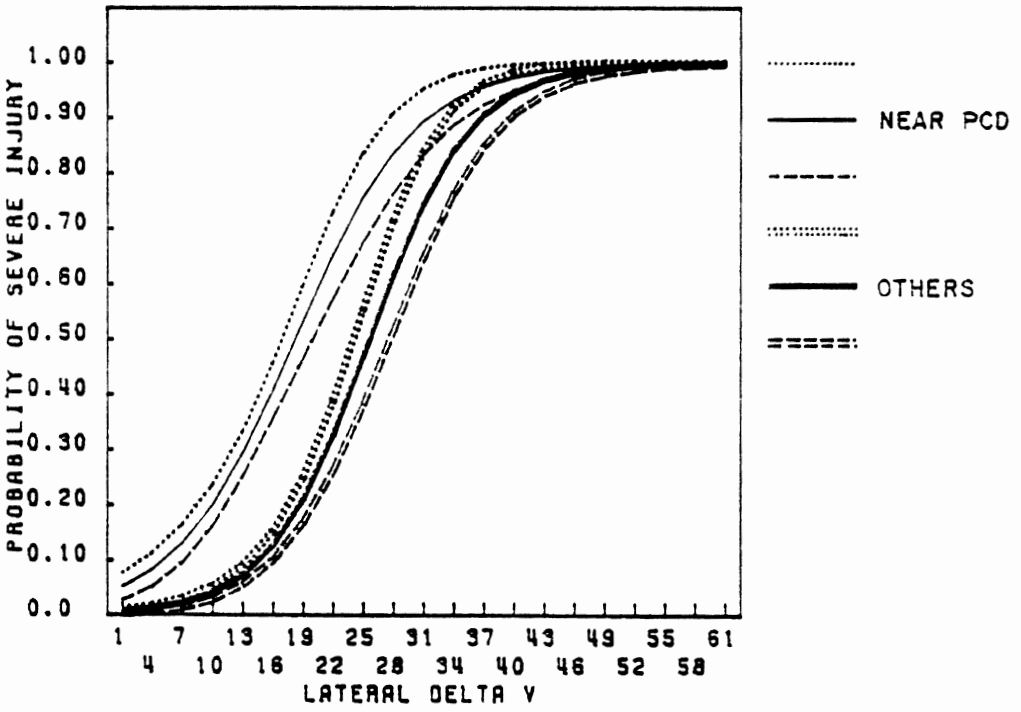

FIGURE 3.56 Logistic Curves of Two-Variable Models (Lateral Delta V, Age) For Near PCD and (Far Occ + Near NPCD) Phases 1 and 2 - Side Impacts

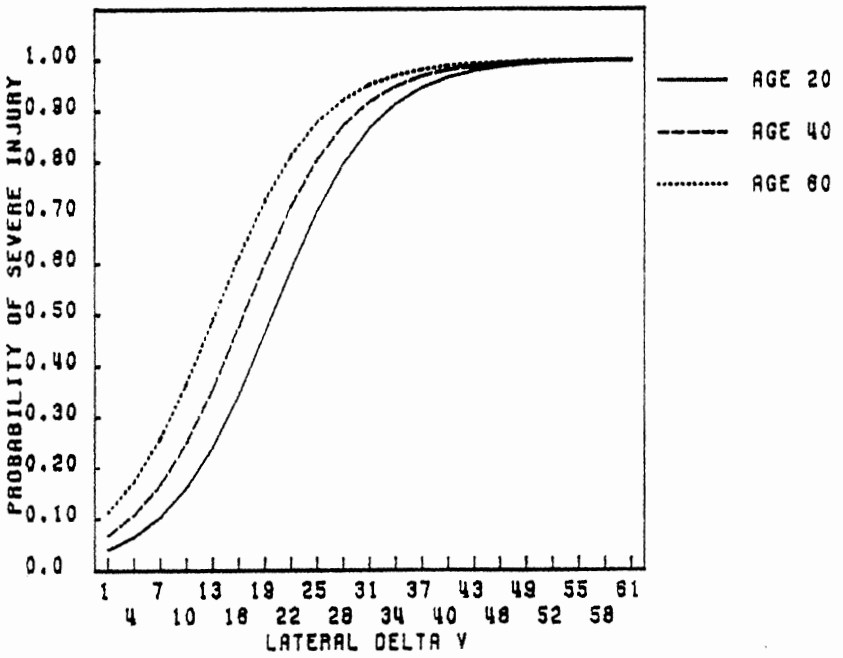

FIGURE 3.57 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For Near PCD Phases 1 and 2 - Side Impacts

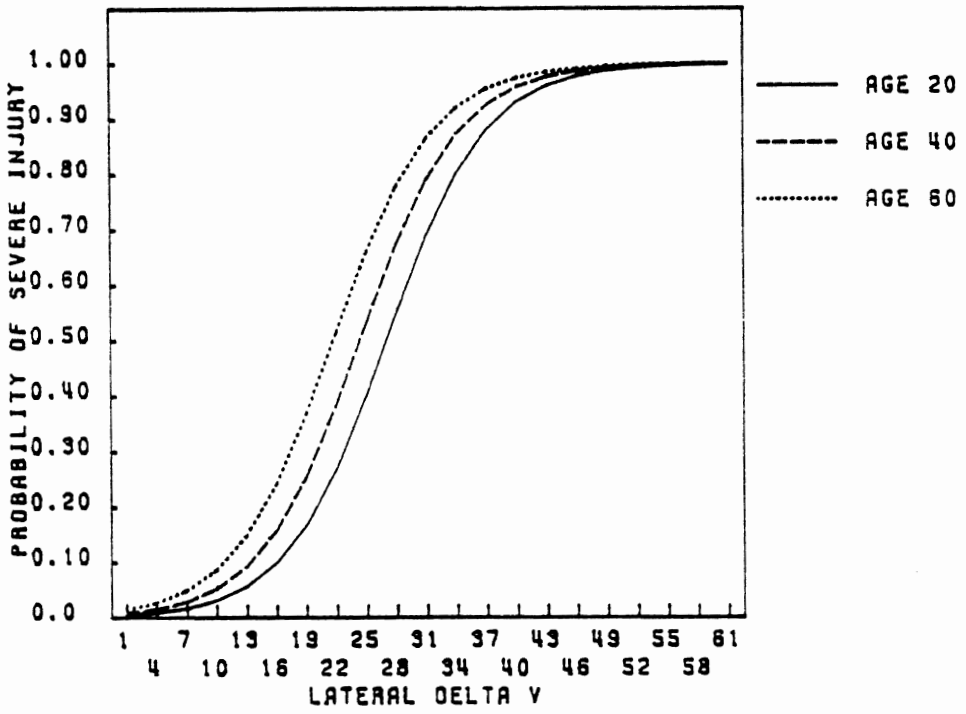

FIGURE 3.58 The Age Effect of Two-Variable Model (Lateral Delta V, Age) For (Far Occ + Near NPCD) Phases 1 and 2 - Side Impacts

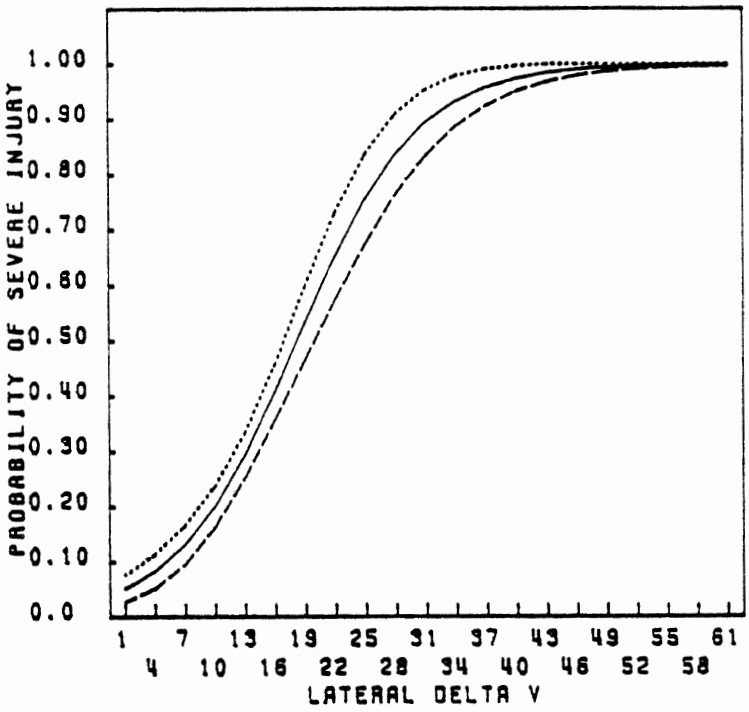

FIGURE 3.59 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For Near PCD Phases 1 and 2 - Side Impacts

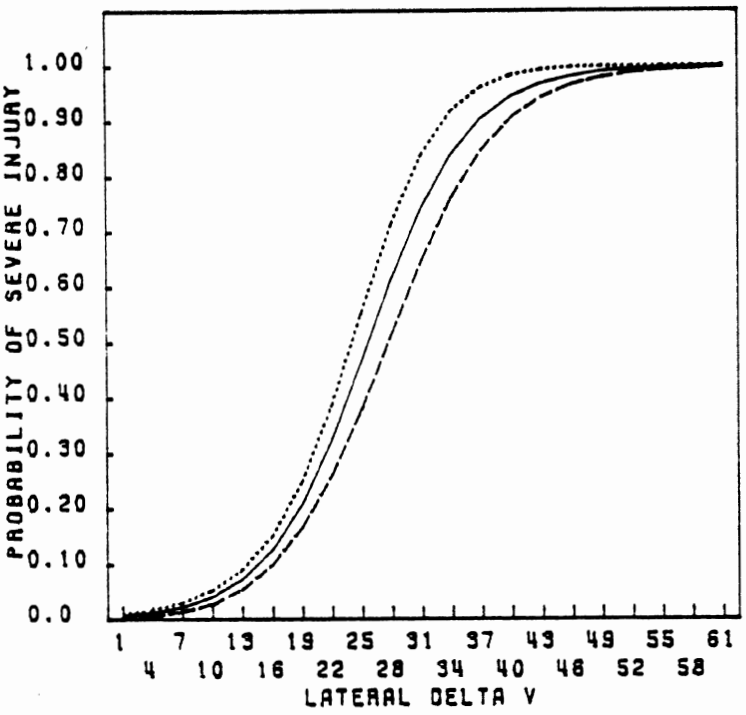

FIGURE 3.60 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Lateral Delta V, Age) at Age 30 For (Far Occ + Near NPCD) Phases 1 and 2 - Side Impacts

3.3.4 Model Estimation - Phases 1 and 2 Combined. In order to improve the predictive capability of Equation 3-57 and 3-58, particularly in predicting severe injuries, other potential independent variables were investigated. They were:

Vehicle Weight
Object Contacted
Rural/Urban
Vehicle Model Year
Ejection
Restraint Usage
Contact Point
Body Region
Injury Type

Each of these variables was examined with a view to introducing it into the model where Lateral Delta V and Age were present. Results from these analyses are given below.

Object Contacted. Object contacted was incorporated into the modelling as a dummy variable. This variable was coded 0 if the object contacted was a passenger car and 1 otherwise. This dummy variable, when incorporated into the model in the presence of Lateral Delta V and Age, was found to be significant in the Near PCD subset but not in the other subset. The estimated model is as follows:

Estimated Model with Lateral Delta V, Age and Object Contacted

Near PCD (N=650, LRS=171.87, DF=3)

(3-59) $\hat{p}_i = F(2.2692 - 0.0919X_1 - 0.0154X_2 - 0.6458X_3)$

where

$\hat{p}_i$ is the probability of a non-severe injury,
F is the logistic distribution,
X_1 is Lateral Delta V,
X_2 is Age,
X_3 is 0 if Object Contacted was a passenger car and 1 otherwise.

The goodness of fit of Equation 3-59 is as follows:

Overall percent correct prediction is 75.7%

Percent correct prediction of non-severe injuries is 90.1%

Percent correct prediction of severe injuries is 46.8%

It can be seen that object contacted did improve the predictive capability of the Near PCD model, particularly it improved the predicted severe injuries by 5% (or 11 cases). Severe injuries had been the target for the improvement.

Rural/Urban. Rural/Urban was incorporated into the modelling in the presence of Lateral Delta V and Age. It did not appear to be a statistically significant variable; nor did it improve the models' predictive capability.

Vehicle-Year. Vehicle-Year was recoded as a 3 level variable with the levels defined as pre-1968, 1968-1973, and 1974-1978. This three level categorical variable was incorporated into the model as 2 dummy variables. It was found that Vehicle-Year did not appear to be a significant independent variable.

Ejection. The majority of occupants (about 90% or more) were not associated with either ejection or entrapment. Table 3.25 shows the number of occupants ejected and otherwise for cases with valid NEWOAIS3 Code, Lateral Delta V and Age; the proportions of severe injuries are also shown.

TABLE 3.25

Number of Ejection with Valid
OAIS Code, Lateral Delta V and Age

Ejection	Near PCD		Far Occ. + Near NPCD	
	Number of Cases	Percentage of Severe Injuries (%)	Number of Cases	Percentage of Severe Injuries (%)
No Ejection	563	28.2	1217	7.7
Ejection	32	43.8	25	36.0
Entrapment	23	73.9	6	50.0

Table 3.25 indicated that Entrapment in general would result in the highest proportion of a severe injury (50% or over), Ejection would result in a higher probability of a severe injury (slightly lower than 50%) than when No Ejection was involved. Near-side occupants appeared to have higher rates of Entrapment and Ejection than far-side occupants; their ejection/entrapment was also more likely to result in a severe injury.

Ejection, however, when incorporated into the modelling in the presence of Lateral Delta V and Age, did not appear to be a significant variable. This could be attributable to the much smaller sample sizes of cases with Entrapment or Ejection and/or the fact that Ejection might be correlated with Lateral Delta V.

Restraint Usage. The majority of occupants were found not using any kind of restraint devices. Only a handful used Torso and Lap restraints or Lap-Only restraints. Table 3.26 shows the number of occupants with and without restraint for cases with valid OAIS Codes, Lateral Delta V and Age and the associated proportion of severe injuries. For Near PCD, an occupant without a restraint or not using restraint appeared to be more prone to a severe injury than when some kind of restraint was used. This, however, was not obvious with the (Far Occ + Near NPCD) subset.

Restraint usage, when brought into the modelling in the presence of Lateral Delta V and Age, did not appear to be statistically significant.

In summary, the model estimation results are described by Equations 3-58 and 3-59. The Near PCD the model contained Lateral Delta V, Age and Object Contacted. The model for the Far Occupants + Near NPCD subset included only Lateral Delta V and Age.

TABLE 3.26

Number of Occupants With and Without Restraints

Phases 1 and 2 - Side Impacts

	Near PCD			Far Occ + Near NPCD		
	Number of Cases	Percentage of Severe Injuries	Mean Lateral Delta V	Number of Cases	Percentage of Severe Injuries	Mean Lateral Delta V
Not Used	568	33.6	12.9	1108	10.3	10.9
Lap/Torso	16	25.0	12.1	41	9.8	9.2
Lap Only	12	8.3	9.0	40	5.0	9.3
Others	1	0	-	4	25.0	-
No Restraint	23	43.5	13.1	38	10.5	11.4

3.3.5 Model Evaluation - Phase 1 and Phase 2 Combined. Tables 3.27 and 3.28 list the injury types and the body regions which were associated with relatively high proportions of severe injuries to total injuries. These severe injuries were frequently being mispredicted by the models represented by Equations 3-58 and 3-59. Tables 3.29 and 3.30 cross-tabulate the injury types and the affected body regions associated with the injuries which the established models predicted incorrectly more frequently than not. Only the combinations of Injury Types and body region which resulted in high proportions of severe injuries to total injuries were included in the tables. For each of these combinations, the associated values of Lateral Delta V, Age and the Contact Point were also indicated.

The injury types which had large proportions of severe injuries and large proportions of severe injuries being mispredicted by the models were Rupture, Dislocation and Fracture. The body regions which often suffered severe injuries and the severe injuries of which were frequently mispredicted by the models were Abdomen, Chest, and, to a lesser extent, Forearm, Pelvic/Hip and the Lower Limbs. These cases were all associated with low to moderate values of Lateral Delta V (and therefore low to moderate Delta V). Correct prediction of these cases would require variables in addition to Lateral Delta V, Age and Object Contacted.

Tables 3.29 and 3.30 indicated the various combinations of body regions and the subsequent injury types in which further investigation would be required if the models' predictive capability were to be considerably improved. Examination of such cases revealed that

1. All cases had low to moderate values of Lateral Delta V.

2. The majority of these cases involved the occupants coming into contact with the side interior of the vehicles and/or the steering assembly. Other contact points tended to be much less common although not insignificant. For example, there were two cases in which the near-side occupants came into contact with the A-pillar with Lateral Delta V of only 3 to 5 mph and both resulted in severe head/skull injuries.

It seemed intuitively reasonable from Tables 3-29 and 3-30 to expect that Contact Points coupled with Body Region might enhance the

prediction of injury severity in the presence of Lateral Delta V and Age.

TABLE 3.27

List of Injury Types with Large Proportions of Severe Injuries and Misprediction

Phases 1 and 2 - Side Impacts

Subset	Injury Type*	Total Number of Cases	Number of Severe Injuries	Number of Mispredicted Severe Injuries
Near PCD	Rupture	7	7	5
	Crushing	6	6	2
	Dislocation	9	8	5
	Hemorrhage	4	3	2
	Fracture	160	102	64
Far Occ + Near NPCD	Rupture	7	7	2
	Fracture	119	65	56
	Dislocation	7	3	3

*Injury Type was ranked by the larger proportion of severe injuries.

Contact Point. For cases with valid Delta V and Age, the majority of occupants for Near PCD were linked with the side-interior (37%), window-glass (11%), steering (9%), armrest (9%) and no-contact (8%). The proportions of severe injuries to total injuries resulting from these contact points were:

side-interior	57%
window-glass	8%
steering assembly	48%
armrest	70%
no-contact	23%

For Far Occ + Near NPCD, the majority of the occupants were found to be linked with side-interior (13%), front-panel (14%), steering (13%), windshield (8%), mirror (7%), window-glass (8%), front-seatback

TABLE 3.28

List of Body Region with Large Proportions
of Severe Injuries and Misprediction

Phases 1 and 2 - Side Impacts

Subset	Body Region*	Total Number of Cases	Number of Severe Injuries	Number of Mispredicted Severe Injuries
Near PCD	Abdomen	32	30	23
	Chest	98	70	32
	Forearm	12	9	6
	Low Ext,Low Leg			
	Ankle/Foot	17	8	5
	Pelvic/Hip			
	and Thigh	47	18	15
	Head/Skull	92	24	9
Far Occ + Near NPCD	Abdomen	23	12	8
	Chest	96	49	38

*Subject was ranked by the larger proportion of the severe injuries

(5%) and no-contact (12%). The proportions of severe injuries to total injuries resulted from these contact points were:

side-interior	22%
front-panel	20%
steering	33%
windshield	6%
mirror	0%
window-glass	0%
front-seatback	26%
no-contact	13%

The Contact Point variable contained a considerable number of missing data. When this variable was incorporated into the modelling in the presence of Lateral Delta V and Age, the number of valid cases was reduced from 650 to 326 for Near PCD and from 1293 to 387 for Far Occ + Near NPCD. Moreover, the number of severe injuries for Near PCD was reduced from 216 to 135 and for the other subset from 131 to 64 cases.

TABLE 3.29

Combination of Injury Type and Body Region with Large Proportions of Severe Injury and Misprediction for Near PCD

Phases 1 and 2 - Side Impacts

Injury Type	Body Region	Percent[1] Severe Injury	Percent[2] Severe Injury Mis-predicted	Number of Severe Injuries	Delta V of Mis-predicted Cases	Age of Mis-predicted Cases	Contact Point
Rupture	Abdomen	100	71	7	6-12	18-30	side interior[3], steering
Dislocation	Neck	100	50	4	5-19	13-17	no contact
	Pelvic/Hip	100	100	1	14	21	side interior
Fracture	Neck	100	67	9	8-16	16-55	no contact
	Chest	78	53	49	4-19	22-84	side interior
	Forearm	90	67	9	9-18	7-57	side interior, steering
	Face	67	83	6	2-15	14-43	side interior, glass
	L.Ext.	100	100	2	9-14	19-21	-
	L.Leg	57	50	4	14-15	20-27	side interior
	Thigh	56	80	5	6-17	9-22	side interior
	Ankle/Foot	50	100	1	12	40	side interior
	Pelvic/Hip	44	82	11	11-19	15-55	side interior
Hemorrhage	Abdomen	100	100	1	20	18	side interior
	Chest	100	100	1	12	53	-
Crushing	Head/Skull	100	67	3	3-5	16	A-pillar
Laceration	Abdomen	100	50	6	12-15	16-62	side interior, steering
Contusion	Abdomen	88	93	14	12-20	11-34	side interior, front panel

[1]Percent severe injury is defined as a ratio of severe injuries to total injuries.

[2]Percent severe injury mispredicted is defined as a ratio of the severe injuries that are mispredicted by the models to total severe injuries.

[3]Side Interior includes side interior and armrests

TABLE 3.30

Combination of Injury Type and Body Region with Large Proportions of Severe Injury and Misprediction for Far Occ + Near NPCD

Phases 1 and 2 - Side Impacts

Injury Type	Body Region	Percent[1] Severe Injury	Percent[2] Severe Injury Mis-predicted	Number of Severe Injuries	Delta V of Mis-predicted Cases	Age of Mis-predicted Cases	Contact Points
Rupture	Abdomen	100	33	6	8-25	25-29	side interior[3]
Dislocation	Neck	100	100	1	10	56	no contact
	Elbow	100	100	1	8	40	-
	Pelvic/Hip	100	100	1	18	46	-
Fracture	Neck	90	67	9	12-25	23-57	no contact
	Chest	80	88	40	4-25	8-94	side interior, front panel steering
	Shoulder	38	100	5	5-15	16-77	side interior, front panel
Laceration	Abdomen	100	100	2	11-21	29-51	side interior
Contusion	Abdomen	40	100	4	13-24	19-44	steering

[1]Percent severe injury is defined as a ratio of severe injuries to total injuries.

[2]Percent severe injury mispredicted is defined as a ratio of the severe injuries that are mispredicted by the models to total severe injuries.

[3]Side interior includes side interior and armrests.

Contact Point was also found to be somewhat correlated with Body Region. This only allowed either Contact Point or Body Region to be included in the models but not both. Tables 3.29 and 3.30 suggested that Body Region was likely to be a better independent variable than Contact Point because a body region seemed to immediately suggest a high probability of a certain contact point. Furthermore, Contact Point had the larger number of missing data.

Body Region. For cases with valid Lateral Delta V and Age of the Near PCD subset, most body regions (except head/skull and face) came into contact essentially with only one or at most two contact points, namely, the side interior, the side interior and the armrest, or the side interior and the steering. Neck injuries were mostly "no-contact" injuries.

For Far Occ + Near NPCD, the distribution of Body Region by Contact Point was slightly different from Near PCD. Side Interior, Front Panel and Steering were the more common contact points recorded. Neck and Back injuries were essentially "no-contact" injuries.

3.3.6 Final Models - Phase 1 and Phase 2 Combined. Based on the distribution of Contact Point of each Body Region and the proportion of severe injuries to total injuries of each Body Region, a set of dummy variables was created for Body Region as follows. The Body Region variable was first grouped into six classes:

Head/Skull and Neck
Upper Extremities, Elbow and Forearm
Chest and Abdomen
Pelvic/Hip, Lower Extremities and Thigh
Lower Leg and Ankle/Foot
Others including Missing Data

The above classes of Body Region were then coded as six (0,1) dummy variables. The first five dummy variable were incorporated into the model. When Body Region was incorporated into the modelling it was found that the percent correct prediction of the severe injuries of both Near PCD and Far Occ + Near NPCD increased considerably. For both subsets Body Region showed statistically significant coefficients. The models are shown below:

Final Estimated Models for Phases 1 and 2

Near PCD (N=649, LRS=306.25, DF=8)

(3-60) $\hat{p}_i = F(2.6498 - 0.0905X_1 - 0.0041X_2 - 0.9043X_3 - 0.5492X_4 - {}^{1}1.1180X_5 - 1.8471X_6 - {}^{1}0.9347X_7 - {}^{2}0.8553X_8)^3$

Far Occ + Near NPCD (N=1291, LRS=341.63, DF=7)

(3-61) $\hat{p}_i = F(3.4437 - 0.1001X_1 - 0.0093X_2 - 0.5422X_4 - 0.4324X_5 - {}^{1}1.5676X_6 - 0.7767X_7 - {}^{1}0.2132X_8)^2$

where

$\hat{p}_i$ is the probability of a non-severe injury,

F is the logistic distribution,

X_1 is Lateral Delta V

X_2 is Age,

X_3 is 0 if Object Contacted was a passenger car and 1 otherwise,

X_4 is 1 if the head or neck was injured and 0 otherwise,

X_5 is 1 if the upper extremities, elbow or forearm was injured and 0 otherwise,

X_6 is 1 if the chest or abdomen was injured and 0 otherwise,

X_7 is 1 if the pelvic/hip, thigh, or lower extremities were injured and 0 otherwise, and

X_8 is 1 if the lower leg or ankle/foot was injured and 0 otherwise.

The goodness of fit results of these 2 models are shown in Table 3.31

Figures 3.61 and 3.62 show the histograms of the $\hat{p}_i$ values of both models. The inclusion of Body Region in the Near PCD model in the presence of Lateral Delta V, Age and Object Contacted markedly improved the model's predictive capability. The percent correct prediction of severe injuries increased from 46.8% to 66.5% and the overall correct prediction of both non-severe and severe injuries increased from 75.7% too 81.3%. For (Far Occ + Near NPCD) the percent correct prediction of severe injuries increased from 30.5% to 42.7% and the overall correct

TABLE 3.31

Goodness of Fit

Phases 1 and 2 - Side Impacts

Subset	Model	Sample Size		Percent Correct Prediction		
		Non-Severe	Severe	Overall	Non-Severe	Severe
Near PCD	$\hat{p}_i$ = F(Lateral Delta V Age, Body Region, Object Contacted)	434	215	81.3	88.7	66.5
Far Occ + Near NPCD	$\hat{p}_i$ = F(Lateral Delta V, Age, Body Region)	1160	131	92.3	97.9	42.7

prediction of both non-severe and severe injuries increased from 91.7% to 92.3%.

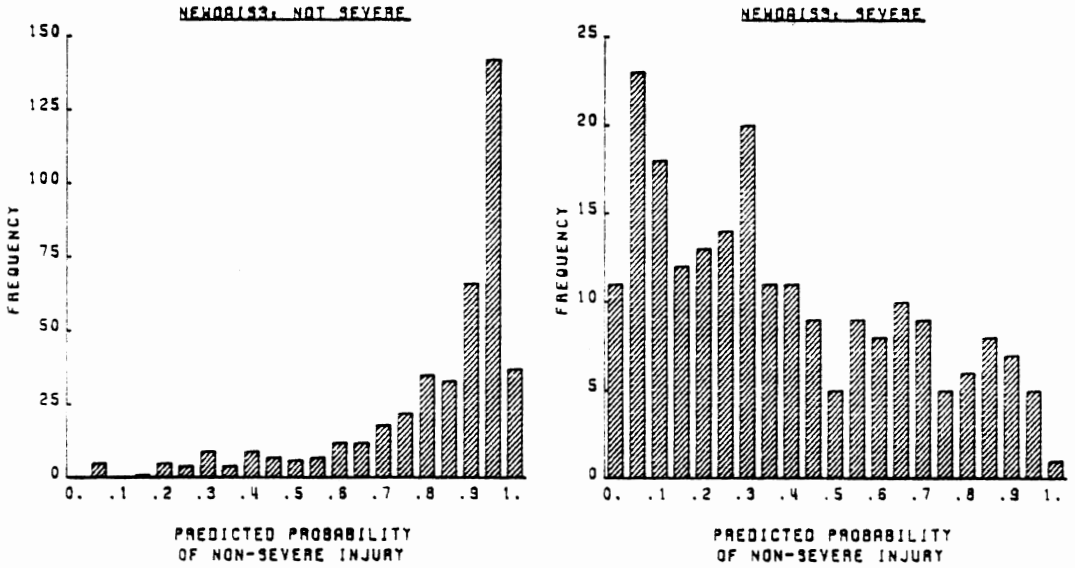

FIGURE 3.61 Histograms of $\hat{p}_i$ of Four-Variable Model (Lateral Delta V, Age, Object Contacted, Body Region) For Near PCD Phases 1 and 2 - Side Impacts

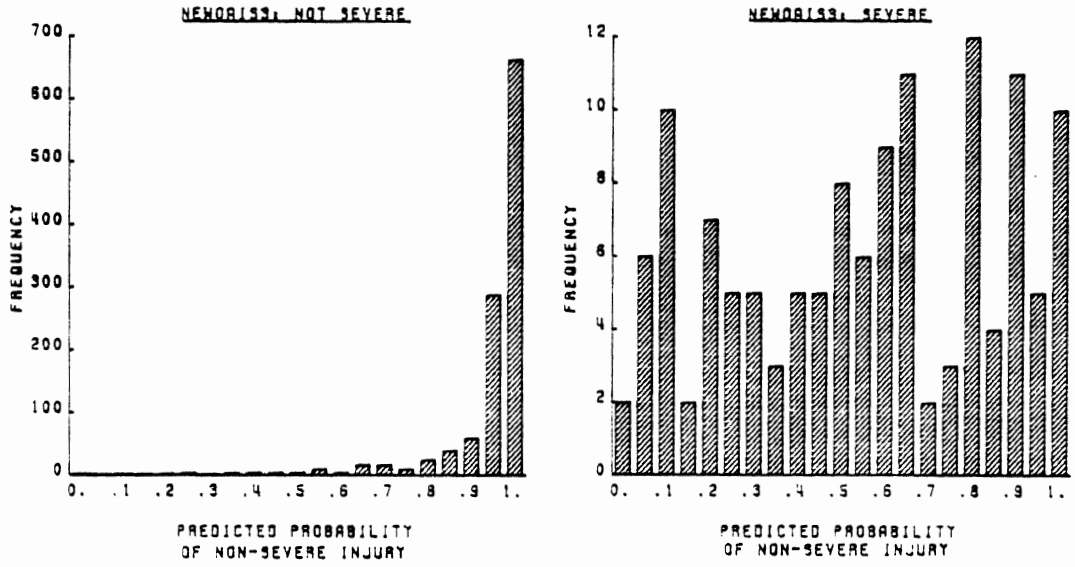

FIGURE 3.62 Histograms of $\hat{p}_i$ of Four-Variable Model (Lateral Delta V, Age, Object Contacted, Body Region) For (Far Occ + Near NPCD) Phases 1 and 2 - Side Impacts

Figures 3.63 shows the estimated confidence limits for the final Near PCD model. Where X_5 is given a value of 1 (i.e., an injury to Upper Extremities or Elbow or Forearm) and X_4, X_6, X_7, X_8 are zero. The three curves represent, when an injury is to Upper Extremities or Elbow or Forearm, and the estimated probability of a severe injury ($1-\hat{p}_i$) as a function of Lateral Delta V fixing Age at 30. The six logistic curves

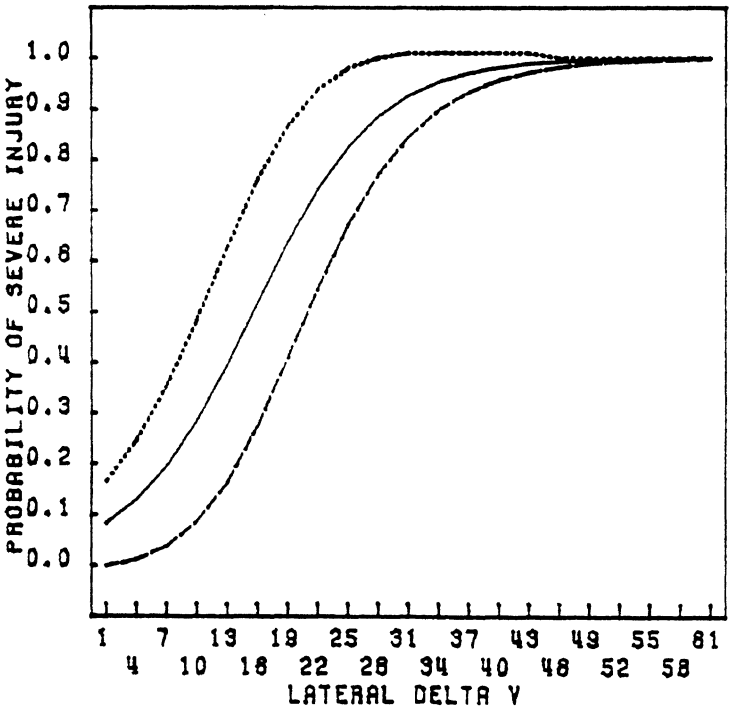

FIGURE 3.63 Confidence Interval of $\hat{p}_i$ of Four-Variable Model (Lateral Delta V, Age, Object Contacted, Body Region) at Age 30 For Near PCD Phases 1 and 2 - Side Impacts

in Figure 3.64 shows the effects of the five levels of the Body Region dummy variable. The six curves are plots of Equation 3-60 to represent the following situations:

- .when the suffered body region was either head/skull or neck
- .when the suffered body region was either upper extremities, elbow or forearm
- .when the suffered body region was either chest or abdomen
- .when the suffered body region was either pelvic/hip, thigh or lower extremities,
- .when the suffered body region was either ankle/foot or lower leg, and
- .when the body region was none of the above or no body region was considered.

The figure shows that Chest and Abdomen are far more likely to sustain a severe injury than any other body regions even at Delta V as low as 10 mph. Injuries in Upper Extremities, Elbow, Forearm, Pelvic/Hip, Lower Extremities, Thigh, Ankle/Foot and Lower Leg are comparable in that they are likely to be severe for Delta V exceeding 20 mph. Head and Neck are likely to sustain a severe injury when Delta V exceeds 22 mph. When

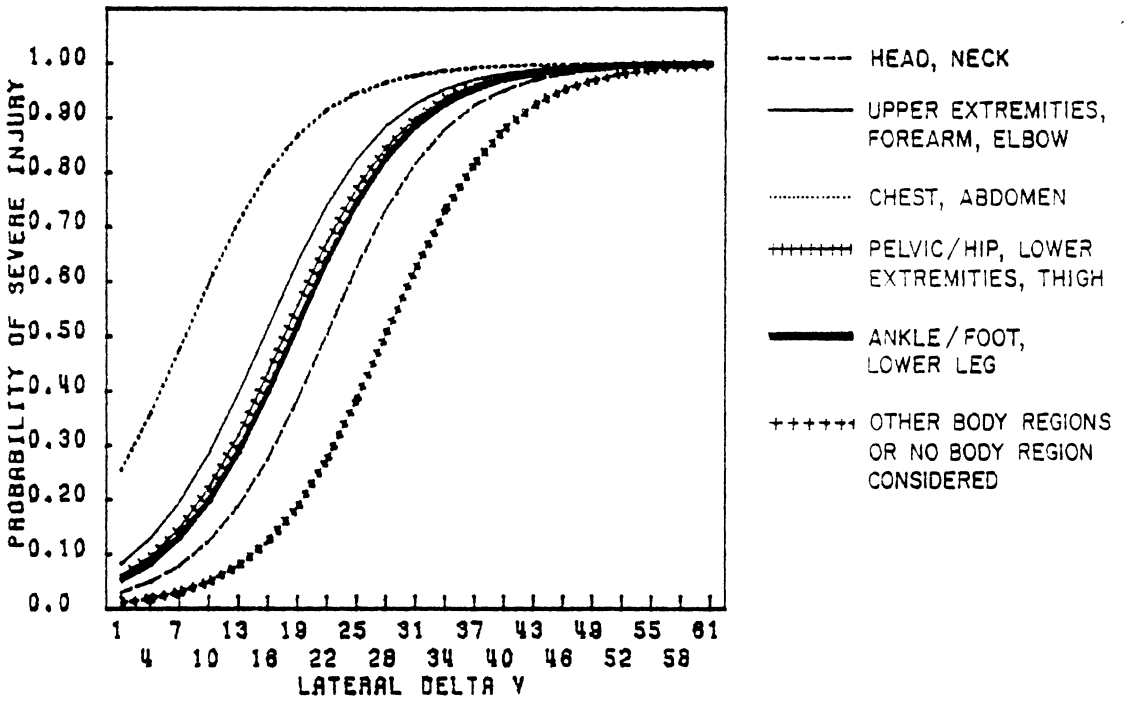

FIGURE 3.64 The Effect of Five Levels of Body Region of Four-Variable Model (Lateral Delta V, Age, Object Contacted, Body Region) For Near PCD Phases 1 and 2 - Side Impacts

injuries are in other body regions not mentioned above or when no body region are specified, they are likely to be severe when Delta V exceeds 27 mph.

Figure 3.65 shows the estimated confidence levels for the final (Occ + Near NPCD) model for the situation where the injury is to Upper Extremities or Elbow or Forearm. Figure 3.66 shows the effects of the five levels of the Body Region dummy variable to represent

.when the injury was to head/skull or neck
.when the injury was to upper extremities, elbow or forearm
.when the injury was to chest or abdomen
.when the injury was pelvic/hip or thigh or lower extremities
.when the injury was to ankle/foot or lower leg
.when the injury was to none of the above or when no body region was considered.

The figure indicates that an injury to either chest or abdomen were far more likely to result in a severe injury that that to other body regions even at Delta V as low as 15 mph. Pelvic/Hip, Lower Extremities and Thigh were susceptible to a severe injury when Delta V exceeds 23

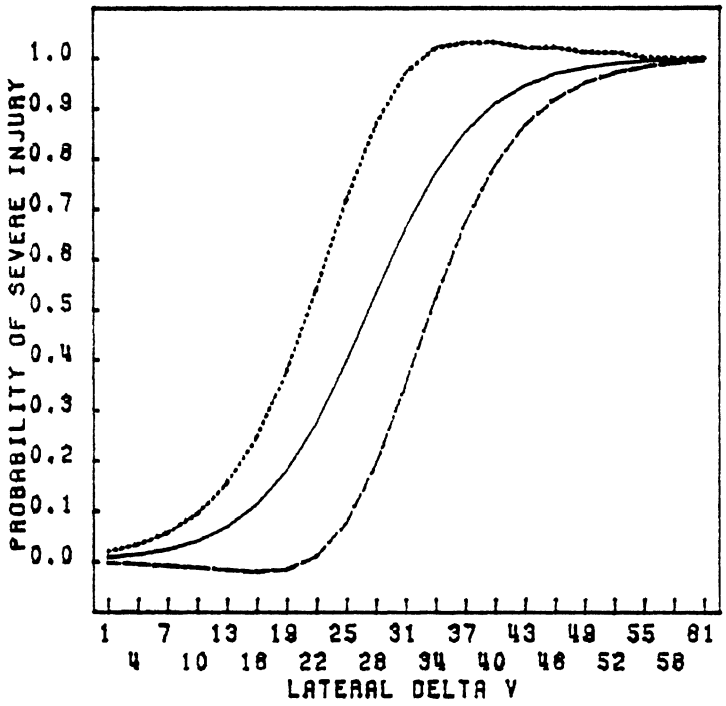

FIGURE 3.65 Confidence Interval of $\hat{p}_i$ of Three-Variable Model (Lateral Delta V, Age, Body Region) at Age 30 For (Far Occ. + Near NPCD) Phases 1 and 2 - Side Impacts

mph. Delta V greater than 25 mph is likely to produce a severe injury in Head or Neck or Upper Extremities or Elbow or Forearm. Ankle/Foot and Lower Leg are likely to sustain a severe injury as Delta V exceeds 28 mph. Injuries to other body regions or where no body regions are specified are likely to be severe as Delta V exceeds 31 mph.

3.3.7 Significant Results. The analysis of the Phase 1 and the Phase 2 data suggested that for the purpose of model estimation the data from both phases should be combined. The model estimation and the model evaluation results also suggested the partitioning of the combined Phases 1 and 2 data into two homogeneous groups (or subsets). These are:

1. those occupants on the same side as the impact which produced passenger compartment damage (Near PCD), and

2. other occupants which includes all occupants on the opposite side to the impact plus those on the same side as the impact which produced no passenger compartment damage (Far Occ + Near NPCD).

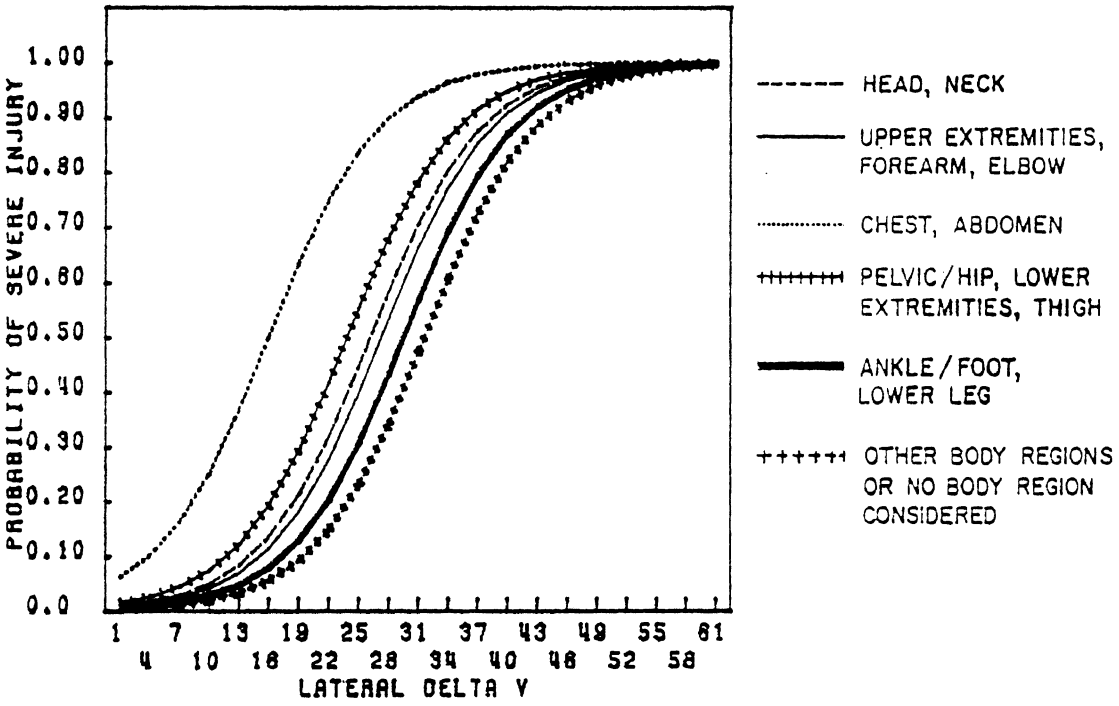

FIGURE 3.66 The Effect of Five Levels of Body Region of Three-Variable Model (Lateral Delta V, Age, Body Region) For (Far Occ + Near NPCD) Phases 1 and 2 - Side Impacts

Earlier modelling work revealed that Lateral Delta V (or the lateral compartment of Total Delta V) was found to be the most single important independent variable in predicting the individual injury severity. Occupant Age was another factor affecting the injury severity. The estimated logit models representing the heuristic relationship between injury severity (non-severe) and Lateral Delta V and Age are as follows:

Near PCD

(3-57) $\hat{p}_i = F(2.1426 - 0.0926X_1 - 0.0152X_2)$

Far Occ + Near NPCD

(3-58) $\hat{p}_i = F(3.2672 - 0.1104X_1 - 0.0146X_2)$

where
$\hat{p}_i$ is the probability of a non-severe injury,
F is the logistic distribution,
X_1 is Lateral Delta V, and
X_2 is Age.

Equations 3-57 and 3-58 state that the estimated probability of a non-severe injury (NEWOAIS3 = 0), $\hat{p}_i$, is a function of Lateral Delta V and Age alone. This probability is bounded by a value of zero and one. A $\hat{p}_i$ value of one predicts that the injury will certainly be non-severe and a $\hat{p}_i$ value of zero predicts that the injury will certainly be severe. A $\hat{p}_i$ value greater than 0.5 will predict a non-severe injury while a $\hat{p}_i$ less than 0.5 will predict a severe injury. Interpretation of the models may be easier if Equations 3-57 and 3-58 are restated in terms of the probability of a severe injury, which is simply $(1-\hat{p}_i)$. The models imply the following:

1. As Lateral Delta V increases, so does the probability of a severe injury. For an occupant age 30 in Near PCD an increase in Lateral Delta V from 15 to 25 mph will take the probability of a severe injury from 0.37 to 0.76.

2. As Age increases, so does the probability of a severe injury. For Near PCD, with Lateral Delta V of 25 mph, a 20-year-old occupant will have the probability of a severe injury of 0.69 and a 40-year-old occupant the probability of 0.81.

3. An occupant in the Near PCD subset is more likely to receive a severe injury than that in the Far Occ + Near NPCD subset given a Lateral Delta V value. For example, for Lateral Delta V of 25 mph and a 30 year old occupant, the probability of a severe injury is 0.76 for an occupant of the Near PCD subset and 0.48 for the Far Occ + Near NPCD subset.

The models as represented by Equations 3-57 and 3-58 predicted the overall injuries fairly well. A closer examination of their predictive capability revealed that when a injury was observed to be non-severe, the models would be correct almost all of the time, but that when an injury was observed to be severe, the models would be correct only 30% to 40% of the time. Although non-severe injuries occurred much more frequently than severe injuries, it is more crucial to be able to reasonably predict the injuries with high severity. One of the reasons for the model to preform somewhat unsatisfactorily in cases of severe injuries is that severe injuries occurred at all ranges of Delta V values, depending upon numerous other factors.

In order to try to improve the predictive capability of the models for a better prediction of severe injuries, a range of other variables were investigated. Restraint Usage and Ejection were examined but the

relatively very small sample size of occupants using restraints or having been ejected or trapped made these variables of little value in modelling. About 90% of the occupants did not use any form of restraint and over 90% were not ejected or trapped in the side impacts.

Object Contacted was investigated and was found significant for the Near PCD subset. This variable implied that for Near PCD occupants, a striking vehicle which was bigger than a passenger car would increase the probability of the injury being severe. Object Contacted was not significant for the Far Occ + Near NPCD subset. Detailed analysis of the mispredicted cases or outliers indicated that injuries involving certain body regions tended to result in severe injuries regardless of Delta V values. There were also certain injury types which occurred at relatively low Delta V values and resulted in severe injuries. Most, if not all, the severe injuries that were being mispredicted by the two-variable models (Later Delta V and Age) had low to moderate Lateral Delta V values and the injuries were the results of certain body regions coming into contact with certain interiors of the vehicle at the impact. The most common contact point(s) seemed to be side interior and/or steering for chest, abdomen, pelvic/hip, forearm and the lower limbs. The less common contact points were A-pillar for head/skull injuries or no contact for neck injuries. It was thought that with Lateral Delta V, Age, Contact Point and Body Region in the models, the occurrence of injury severity would become more explainable. Lateral Delta V is a proxy for the force exerted on the occupant at the impact, causing the occupant to move from his/her original position and causes a body region to come into contact with the interior of the vehicle. The occupant age could give rise to the different resistance to or the tendency for a certain type of injury.

However, the contact point variable did impose some serious problems in modelling. First, the missing data on Contact Point for both severe and non-severe injuries was to the extent that the number of valid cases for modelling would drop by about 50% for severe injuries and over 50% for non-severe injuries. Second, Contact Point showed a correlation with Body Region. It was found that between Contact Point and Body Region, the latter appeared to be a better independent

variable. A body region, with the exception of Head and Face, tended to be associated with either only one or probably two major contact points. A contact point, one the other hand, could imply many different body regions.

That a body region almost immediately implies a certain contact point seemed to justify its inclusion in the model even without another variable such as Contact Point. In fact, the final model for Near PCD with Lateral Delta V, Age Object Contacted, and Body Region as the independent variables showed most considerable improvement in the model's predictive capability; such a model now predicted non-severe injuries correctly 89% of the time and it predicted severe injuries correctly 67% of the time. The Far Occ + Near NPCD model, with Lateral Delta V, Age and Body Region as the independent variables, now correctly predicted non-severe injuries 98% of the time and correctly predicted severe injuries 43% of the time. The influence of the Body Region variable in predicting the injury severity can be illustrated in the following case, where a passenger car was hit by a truck and which involved a Lateral Delta V of 15 mph and an abdominal rupture to a Near PCD occupant age 30. The two-variable model (Lateral Delta V and Age) will predict the probability of a severe injury of 0.37 (i.e., a non-severe injury) while the model with Lateral Delta V, Age, Object contacted and Body Region will predict the probability of severe injury of 0.95 (i.e., a severe injury).

3.4 Preliminary Analytical Results for Frontal Impacts - Phase I Data

This section presents preliminary work to develop mechanistic models for frontal impacts. The analysis reported in this section was carried out on a preliminary version of the data from the first fifteen months of NCSS (January 1977 through March 1978).

The initial modelling efforts were carried out separately for several subsets of the frontal impacted vehicles. The selection of these subsets is described in the first subsection. Examination of candidate independent variables is covered in the second subsection. Modelling results and model evaluation are presented in the third and fourth subsections, while the last subsection summarizes the resulting logit models and their associated confidence limits.

3.4.1 Defining Subsets. Frontal collisions comprise principally two-vehicle accidents and one-vehicle accidents. Multiple vehicle collisions are much less common. The number of cases which involved occupants of towed vehicles with valid NEWOAIS3 coding for these accident categories are shown below:

One-vehicle accidents	1491 cases	(25.5%)
Two-vehicle accidents	3705 cases	(63.4%)
Three-vehicle accidents	537 cases	(9.2%)
Multiple-vehicle accidents	110 cases	(1.9%)

Occupants of towed vehicles (or case vehicles) represented about 90% of all the cases.

For the occupants of case vehicles, the distribution of occupants among the seat positions is shown in Table 3.32. Occupants on the back seats represented about 11% of the total occupants and they were not included in the analysis. For the occupants of case vehicles, the following proportions of occupants by specific Horizontal Location of Deformation (CDC) were observed:

Center plus distributed	2545 (34.8%)	Vehicles
Left plus left-side center	2662 (36.5%)	Vehicles
Right plus right-side center	2092 (28.7%)	Vehicles

Subsetting to obtain homogeneous analytical cells for the modelling was carried out by examining variables such as:

TABLE 3.32

Number of Occupants by Seat Position

Phase 1 Data - Front Impacts

Seat Position	Number of Occupants
Drivers	4488 (62.6%)
Front right	1585 (22.1%)
Front center	305 (4.3%)
Back left	299 (4.2%)
Back right	338 (4.7%)
Back center	155 (2.1%)

(a) Proportion of severe and non-severe injuries (based on NEWOAIS3)[21] to total injuries.
(b) The location of deformation and damage distribution type variables of the CDC
(c) Direction of Force (CDC)
(d) Delta V
(e) Injury Type and Body Region
(f) Seat Position
(g) Number of Vehicles

Ideally, the subsets should be dissimilar as regards the injury severity proportion and the proportion of various injury types from one subset to another, but similar in the ranges and the distributions of the potential independent variables, such as the collision severity variables and the occupant variables. It is also essential for the subsets to contain adequate data to permit reliable modelling results.

Various subsettings were closely examined. The one which, at this stage, appeared most reasonable comprised the following subsets:

1. One-vehicle accidents, distributed and center damages (SHL=D and C) and front-seat occupants; to be referred to as CIA-1VEH (Center Impacts, All Occupants-1 Vehicle).

2. One-vehicle accidents, left and right damages (SHL=L+R+Y+Z) and drivers; to be referred to as OID-1VEH (Off-center Impacts, Drivers only-1 Vehicle).

[21] OAIS refers to the overall abbreviated injury score as defined in The Abbreviated Injury Scale, (1976 Revision), American Association for Automotive Medicine, Norton Grove, Illinois.

3. One-vehicle accidents, left and right damages (SHL=L+R+Y+Z) and front seat passengers; to be referred to as OIP-1VEH (Off-center Impacts, Passengers only-1 Vehicle).

4. Two-vehicle collision (SHL=D and C) and front seat occupants; to be referred to as CIA-2VEH.

5. Two-vehicle collision (SHL=L+R+Y+Z) and drivers; to be referred to as OID-2VEH.

6. Two-vehicle collision (SHL=L+R+Y+Z) and front seat passengers; to be referred to as OIP-2VEH.

The number of cases in each of these six subsets is shown in Table 3.33 together with the number of cases where NEWOAIS3 codes and Delta V are valid.

TABLE 3.33
Number of Cases Valid for Specified Variables

Phase 1 Data - Front Impacts

Subset	Number			Percentage Severe Injuries
	Valid Non-Severe	Valid Severe	Valid Delta V	
CIA-1VEH	289	104	308	26.5
OID-1VEH	514	116	548	18.4
OIP-1VEH	231	43	242	15.7
CIA-2VEH	957	185	1062	16.2
OID-2VEH	1310	154	1247	10.6
OIP-2VEH	532	70	539	12.1

Table 3.33 also shows the proportion of the severe injuries to total injuries for the six subsets. The first three subsets were occupants of single-vehicle accidents while the other three subsets were occupants of occupants of two-vehicle accidents. Occupants of single-vehicle accidents appeared to have a higher proportion of severe injuries than those of two-car accidents. For both single-vehicle and two-vehicle collisions, center impacts were also found to have higher proportions of severe injuries than off-center impacts.

3.4.2 Examination of the Independent Variables. It was envisioned that the variables listed below should be investigated with a view to identifying the influences on the prediction of injury severity:

1. Accident Level Variable
   Rural/Urban

2. Vehicle Level Variables
   Delta V
   Direction of Force (CDC)
   Vertical Location of Deformation (CDC)
   Damage Distribution Type (CDC)
   CDC Extent
   Object Contacted
   Vehicle Weight
   Contact Point
   Intrusion Location

3. Occupant Level Variables
   Age
   Sex
   Height
   Occupant Weight
   Restraint Usage
   Ejection

The following section briefly describes these key independent variables across the six subsets.

Delta V. The range of Delta V for the six subsets are shown in Table 3.34. Figure 3.67 shows the cumulative distribution plots of Delta V for the six subsets. For the three single-vehicle subsets, CIA-1VEH, OID-1VEH and OIP-1VEH, the cumulative curves were close together. For the three two-vehicle subsets, the curve for CIA-2VEH appeared to lie somewhat to the right of those for OID-2VEH and OIP-2VEH. Overall, the six subsets did not show marked differences in the cumulative distributions of Delta V.

Principal Direction of Force. The two-way tables of injury severity (NEWOAIS3) and Principal Direction of Force for the six subsets indicated that for the accidents involving single cars the proportions of severe injuries to total injuries were found to be higher with a twelve-o'clock direction than with other directions of force. For two-vehicle collisions, the proportions of severe injuries to total injuries were higher with a twelve-o'clock, a one-o'clock, and a eleven-

TABLE 3.34
Comparison of Delta V Amongst the Subsets

Phase 1 Data - Front Impacts

Subset	Range	Delta V Mean	S.D.
CIA-1VEH	4-97	18.4	12.3
OID-1VEH	1-75	16.9	10.3
OIP-1VEH	2-68	16.6	10.7
CIA-2VEH	2-99	18.3	10.4
OID-2VEH	2-61	14.0	8.7
OIP-2VEH	2-56	14.6	9.6

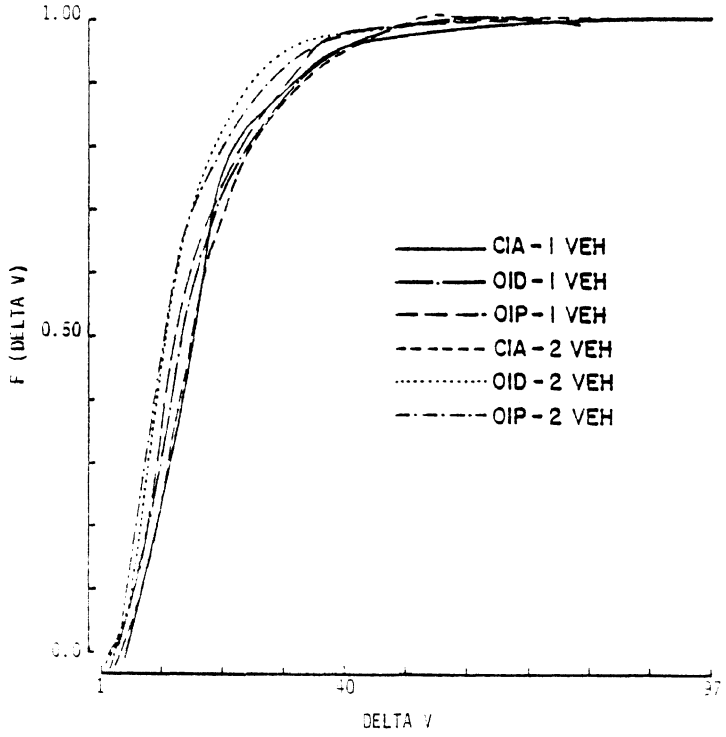

FIGURE 3.67 Cumulative Distributions of Delta V For The Front-Impact Subsets
Phase 1 Data - Front Impacts

o'clock direction than with other directions. When Delta V was held constant, however, the differences appeared much less apparent. Within each direction of force, a higher proportion of severe injuries to total injuries was also noted when vehicles were hit in the center than when

the impact was to the right or to the left. Accidents involving single vehicles displayed higher proportions of severe injuries to total injuries than those involving two vehicles. Table 3.35 summarizes the number of cases by Principal Direction of Force for occupants of one-car and two-car accidents with valid NEWOAIS3 codes.

TABLE 3.35
Number of Occupants by CDC Direction

Phase 1 Data - Front Impacts

	CDC Direction				
	12 O'clock	1 O'clock	11 O'clock	2 O'clock	10 O'clock
One-car accidents	984 (77.9%)	152 (12.0%)	108 (8.6%)	14 (1.0%)	5 (0.4%)
Two-car accidents	1583 (47.5%)	631 (19.0%)	838 (25.2%)	134 (4.0%)	144 (4.3%)

Finally, although there did not appear to be a distinct relationship between Principal Direction of Force and Delta V, it was noted that Delta V's greater than 40 mph were associated with only 12 o'clock, 11 o'clock, and 1 o'clock directions.

Vertical Location of Deformation. For occupants in accidents involving single vehicles and with center impacts, "below-glass" accounted for 91% of the cases and "low" another 8%. For occupants in single-vehicle accidents with off-center impacts, about 93% were "below glass," 3% "low," and 3% "all." For occupants in two-vehicle collisions, about 91% of the cases were "below glass," 6% "middle," and 3% "all." For all subsets proportions of severe injuries to total injuries with the exception of "all" were similar. For "all" the proportion was more than twice as high as the other levels.

Damage Distribution Type. For occupants in single-vehicle accidents with center impacts, 65% of the cases were of "wide"

damage distribution and the other 35% were of "narrow" damage distribution. For occupants in single-vehicle accidents with off-center impacts, 48% were "wide," 27% were "narrow," and another 24% were "corner." For occupants in two-vehicle accidents with center-collisions, 97% of the cases were "wide;" while those with off-center impacts, 77% were "wide" and 20% were "corner."

Proportions of severe injuries to total injuries, with the exception of "corner" for the two-vehicle collisions, were not that much different for all levels. With the "corner" damage in two-vehicle accidents, such proportion appeared much higher.

Object Contacted. For occupants in single-vehicle accidents, the accidents as shown in Table 3.36 were observed.

TABLE 3.36

Object Contacted by Impact Location
for Single-Vehicle Accidents

Phase 1 Data - Front Impacts

Object Contacted	Percentage for Center Impact	Percentage for Off-Center Impact
Utility Poles	24	33
Trees	20	31
Culvert Curbs	12	6
Embankment Abutment	15	6
Guard Rails Bridge Rails	15	8
Non-moving Objects	2	6

For occupants in two-vehicle collisions, the accidents as shown in Table 3.37 were observed.

For occupants in single-vehicle accidents, the proportions of severe injuries to total injuries appeared to be slightly higher for trees, culvert curbs, abutments, embankments and buildings. This was also true when Delta V was held constant. For two-vehicle accidents,

TABLE 3.37

Object Contacted by Impact Location
of Two-Vehicle Accidents

Phase 1 Data - Front Impacts

Object Contacted	Percentage for Center Impact	Percentage for Off-Center Impact
Passenger cars	76	77
Trucks	13	13
Tractor-trailers	3	2
Unknown vehicles	4	4

such proportions were higher when the striking vehicles were larger in size.

Vehicle Weight. The range of vehicle weights and their distributions for all subsets were similar. The variation of the proportions of severe injuries to total injuries with vehicle weights when Delta V was held constant was not apparent for either single vehicle or two vehicle accidents. The scatter plots of Delta V and Vehicle Weights did not reveal an easily detectable relationship.

Rural/Urban. For occupants in single-vehicle accidents, about 60% of the cases were associated with accidents occurring in urban areas. This proportion was much higher (about 75%) for two-vehicle accidents. The proportion of severe injuries to total injuries in the rural and the urban areas for the six subsets are shown in Table 3.38. When Delta V was held constant,the rural accidents also had slightly higher proportions of severe injuries to total injuries than the urban accidents.

Intrusion Location. For occupants in single-vehicle accidents, about 60% of the cases had no intrusion, about 6% had intrusion involving steering column and/or A-pillars, about 3% involving roofs, and about 2% involving sides and about 30% involving other combinations. For two-vehicle accidents, these proportions were found to be somewhat different. No intrusion represented about 75% of the cases,

TABLE 3.38
Proportion of Severe Injuries to Total Injuries by Rural/Urban

Phase 1 Data - Front Impacts

Subset	Rural		Urban	
	Percentage Non-Severe	Percentage Severe	Percentage Non-Severe	Percentage Severe
SINGLE VEHICLE				
CIA-1VEH	71	29	76	24
OID-1VEH	77	23	84	16
OIP-1VEH	78	22	89	11
TWO VEHICLE				
CIA-2VEH	68	32	90	10
OID-2VEH	74	26	94	6
OIP-2VEH	79	21	91	9

intrusion involving steering columns and/or A-pillars was about 5%, those involving roofs was about 1%, and those involving other combinations about 15%.

Proportion of severe injuries to total injuries varied greatly across the locations of intrusion. This proportion was found to be the highest (at least 40%) when the intrusion included steering columns and/or A-pillars; this was followed by the combination of intrusions and roof intrusion. Such proportion was found to be relatively low for cases with no intrusion (5 to 13%).

The scatter plot of Delta V and the various intrusion levels indicated that the Delta V's associated with intrusions of side, side-override and roof tended to be lower than those associated with intrusions of other kinds.

Restraint Usage. In all subsets the majority of the occupants did not use any kind of restraint (about 87%); about 4 to 9% did not have restraint devices; about 3% used lap-and-torso restraints;and about 3% used lap-only restraints; usage of other forms of restraints was relatively rare.

Cursory examination of proportions of severe injuries to total injuries by restraint usage indicated that with the exception of the single vehicle, off-center and drivers subset, the proportion was much lower when lap-torso restraints were used than when no restraints were used or when lap-only restraints were used.

Ejection. The majority of the occupants were associated with no-ejection accidents (well over 90%). It was noted that the proportion of the occupants being ejected and trapped was higher for single-vehicle accidents than for two-vehicle accidents. The proportion of severe injuries to total injuries, in all the subsets,was considerably lower for no-ejection cases than for cases with ejection or entrapment.

Age. There were very little differences in the ranges and the distributions of the Age between occupants in single-vehicle accidents and those in two-vehicle accidents. Small differences did exist between drivers and non-drivers subsets in that there were no drivers under a certain age.

The proportion of severe injuries to total injuries, when Delta V was held constant, indicated that occupants over 30 years old tended to be associated with higher proportions of severe injury.

The scatter plot of Delta V by Age indicated that almost all occupants less than 12 years old and over 65 had Delta V of less than 30 mph, and that occupants between the ages of 18 to 30 years old had the largest range of Delta V values (1 to 98 mph).

3.4.3 Model Estimation. The multivariate logit model described in Section 3.1 was used in the analysis of frontal impacts. Initially, univariate models with Delta V or CDC Extent were tried. The development of multivariate models was described later in this section.

Univariate Models. Univariate models were used to compare the predictive capability of Delta V and CDC Extent in each of the six frontal subsets. Other independent variables, on their own, are not likely to be as good explanatory variables as the above mentioned. They will be significant in explaining injury severity when they are present in the models with either Delta V or CDC extent. Comparisons of

the results of these two models revealed that Delta V was a far better explanatory variable of injury severity than the CDC extent variable (crush measurement). Because Delta V and CDC Extent are highly correlated, the presence of one variable in the model excludes the other. The estimated models for the six subsets with Delta V as the independent variable are shown below and their goodness of fit results in Table 3.39.

TABLE 3.39
Goodness of Fit

Severity = F(Delta V)

Phase 1 Data - Front Impacts

Subset	Sample Size		Percentages Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
CIA-1VEH	172	73	75.5	94.2	31.5
OID-1VEH	346	92	80.1	96.8	17.4
OIP-1VEH	166	28	86.6	98.8	14.3
CIA-2VEH	728	126	89.3	98.4	37.3
OID-2VEH	932	110	92.2	98.5	39.1
OIP-2VEH	371	53	88.9	98.4	22.6

Estimated Models with Delta V

CIA-1VEH (N=245, LRS=39.23, DF=1)

(3-62) $\hat{p}_i = F(1.3595 - 0.0452X_1)$

OID-1VEH (N=438, LRS=63.04, DF=1)

(3-63) $\hat{p}_i = F(1.6351 - 0.0481X_1)$

OIP-1VEH (N=194, LRS=19.44, DF=1)

(3-64) $\hat{p}_i = F(1.7488 - 0.0398X_1)$

CIA-2VEH (N=854, LRS=237.34, DF=1)

(3-65) $\hat{p}_i = F(2.9144 - 0.0881X_1)$

OID-2VEH (N=1042, LRS=262.62, DF=1)

(3-66) $\hat{p}_i = F(2.9683 - 0.0969X_1)$

OIP-2VEH (N=424, LRS=65.85, DF=1)

(3-67) $\hat{p}_i = F(2.1711 - 0.0602X_1)$

where

$\hat{p}_i$ is the estimated probability of a non-severe injury,
F^{-1} is the logistic distribution,
X_1 is Delta V, and
LRS is the Likelihood Ratio Statistic.

Multivariate Models

A large number of multivariate models were tried. At least some mention is made of nearly all the variables tried regardless of whether they were useful or not. The sub-headings indicate the independent variables.

Principal Direction of Force was brought into the modelling in two ways--as an independent variable in the presence of Delta V and by replacing Delta V by Longitudinal Delta V. Neither of these approaches, however, resulted in significant improvement in the predictive capability of the models represented by Equations 3-62 to 3-67.

When Age was incorporated into the model in the presence of Delta V, the following estimation results were obtained for the six subsets:

Estimated Models with Delta V and Age

CIA-1VEH (N=244, LRS=41.54, DF=2)

(3-68) $\hat{p}_i = F(1.6189 - 0.0461X_1 - 0.0084X_2)$

OID-1VEH (N=434, LRS=71.07, DF=2)

(3-69) $\hat{p}_i = F(2.1335 - 0.0522X_1 - 0.0132X_2)$

OIP-1VEH (N=189, LRS=26.6, DF=2)

(3-70) $\hat{p}_i = F(2.4287 - 0.0443X_1 - 0.0220X_2)$

CIA-2VEH (N=849, LRS=253.72, DF=2)

(3-71) $\hat{p}_i = F(3.5671 - 0.0932X_1 - 0.0154X_2)$

OID-2VEH (N=1033, LRS=283.86, DF=2)

(3-72) $\hat{p}_i = F(3.9390 - 0.1076X_1 - 0.0208X_2)$

OIP-2VEH (N=415, LRS=84.92, DF=2)

(3-73) $\hat{p}_i = F(3.0447 - 0.0668X_1 \; 0.0227X_2)$

where

$\hat{p}_i$ is the probability of a non-severe injury,
F is the logistic distribution,
X_1 is Delta V,
X_2 is Age, and
LRS is the Likelihood Ratio Statistic.

The goodness of fit of these models is shown in Table 3.40. By having Age in the models, the predictive capability of the univariate models, particularly in predicting the severe injuries, improved.

Figures 3.68 - 3.73 show the histograms of $\hat{p}_i$ for the six subsets. Each figure, representing a model for a particular subset, consists of a part of histograms, one for non-severe injuries and the other for severe injuries. The axes of both histograms are identical with one axis

TABLE 3.40
Goodness of Fit

Injury Severity = F(Delta V, Age)

Phase 1 Data - Front Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
CIA-1VEH	171	73	75.8	94.2	32.9
OID-1VEH	342	92	80.9	96.5	22.8
OIP-1VEH	161	28	85.7	98.1	14.3
CIA-2VEH	723	126	89.2	97.5	42.1
OID-2VEH	923	110	92.8	98.6	44.5
OIP-2VEH	363	52	88.9	98.1	25.0

representing the estimated probability of a non-severe injury, $\hat{p}_i$, at a 0.05 interval and the other the number of cases with particular values of $\hat{p}_i$. Again, the relatively poor prediction of severe injuries is evident while the prediction of non-severe injuries was very good for all six subsets. The estimated logistic curves for the six subsets are plotted in Figure 3.74 to illustrate how the probability of a severe injury, $1-\hat{p}_i$, varies with Delta V when Age is fixed at 30. Note the marked difference between the single vehicle and two-vehicle subsets. For Delta V less than about 25 to 30 mph, single-vehicle accidents were more likely to result in severe injuries than two-vehicle accidents. For higher Delta V values, however, two-vehicle accidents were more likely to give rise to severe injuries given the same Delta V values. The effect of Age is shown in Figure 3.75 for the OID-1VEH subset and in Figure 3.76 for the OID-2VEH subset. Each of these figures consists of three curves for Age of 20, 40 and 60. The curves show the variation of the estimated probability of a severe injury ($1-\hat{p}_i$) with Delta V. Both figures indicate that in general older occupants are expected to have higher probabilities of severe injuries than younger occupants for any given Delta V value although the effect of Age is most pronounced when Delta V values are about 25 to 30 mph. The age effect also appears

somewhat more pronounced in the two-vehicle subset. Figures 3.77 - 3.80 show the confidence intervals as a function of Delta V for subsets CIA-1VEH, OID-1VEH, CIA-2VEH, and OID-2VEH. In general, the confidence limits for the two-vehicle subsets are much smaller than for the single-vehicle subsets. This implies that if different sets of data were analysed similar modelling results would be more likely to be repeated for the two-vehicle models than for the single-vehicle models. The analyses of the Phase 2 data in Section 3.5 confirmed this. The confidence intervals for the two-vehicle subsets approach zero as Delta V values become either very small or very large; this is not so with the single-vehicle subsets. Finally, the OID-1VEH and OID-2VEH subsets are compared on Figure 3.15. The figure indicates that for high Delta V values (over about 30 mph), the occupants in a two-vehicle accident have higher probability of severe injuries than those in a single-vehicle accident given the same Delta V values. The reverse is true for lower Delta V values (less than 20 mph).

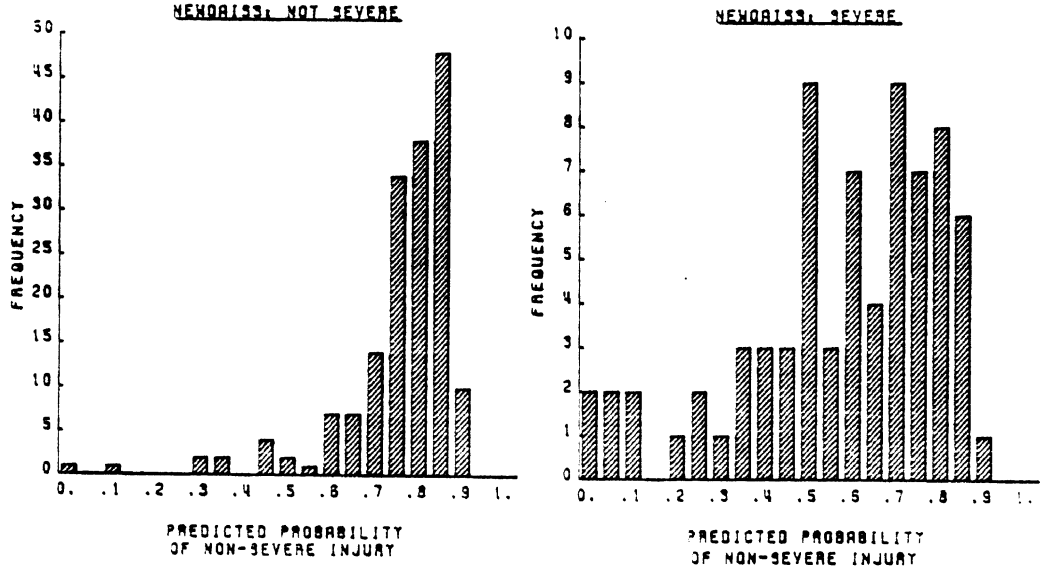

FIGURE 3.68 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For CIA-1VEH Phase 1 Data - Front Impacts

Rural/Urban was incorporated into the model in the form of a dummy variable which gave the value of 1 to rural and 0 to urban. The modelling results indicated that for single vehicle accidents the rural/urban variable was not significant but that it appeared to improve the models' predictive capability for the subsets representing two-vehicle accidents.

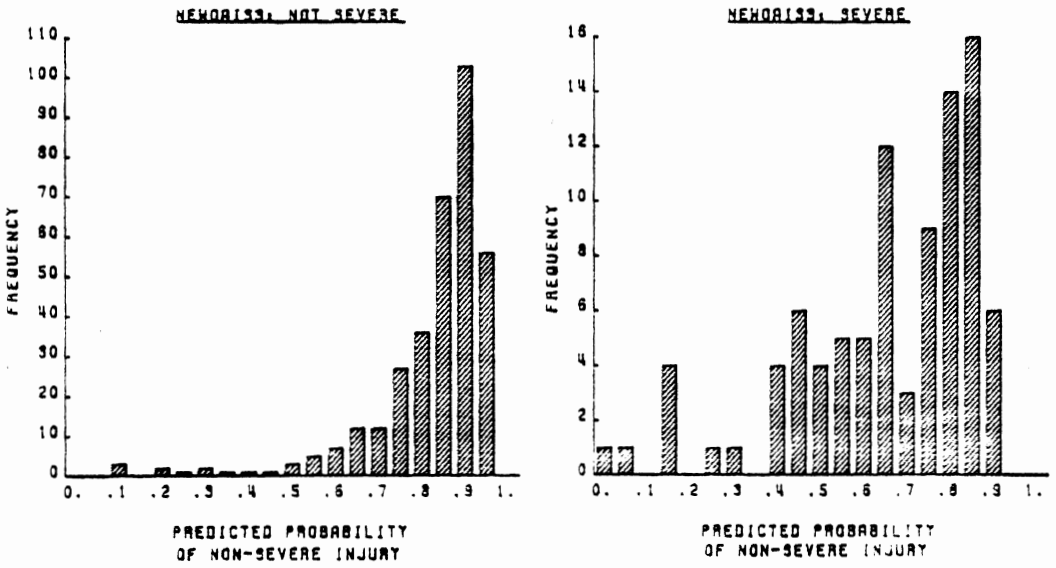

FIGURE 3.69 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OID-1VEH Phase 1 Data - Front Impacts

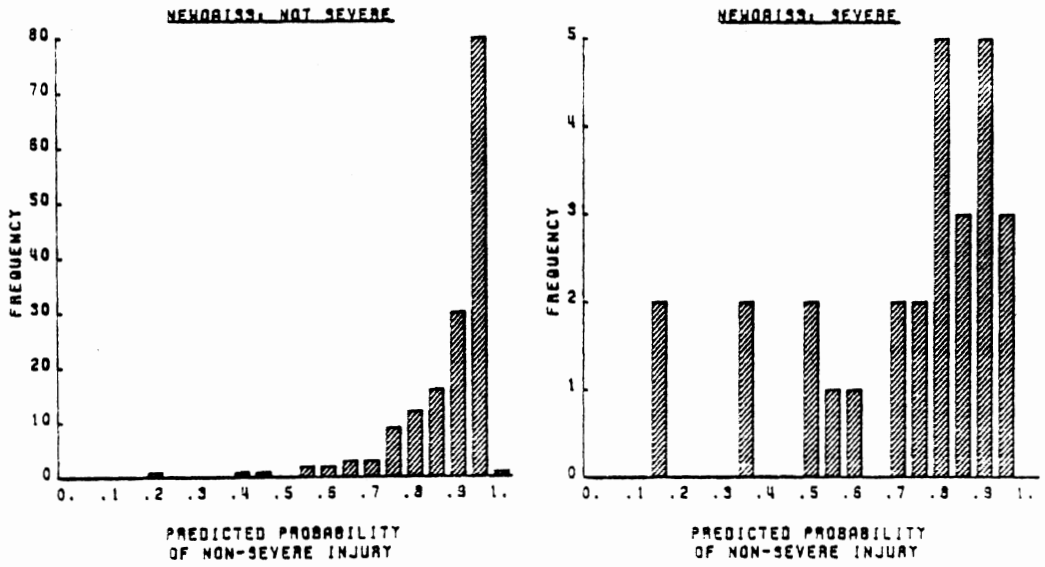

FIGURE 3.70 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OIP-1VEH Phase 1 Data - Front Impacts

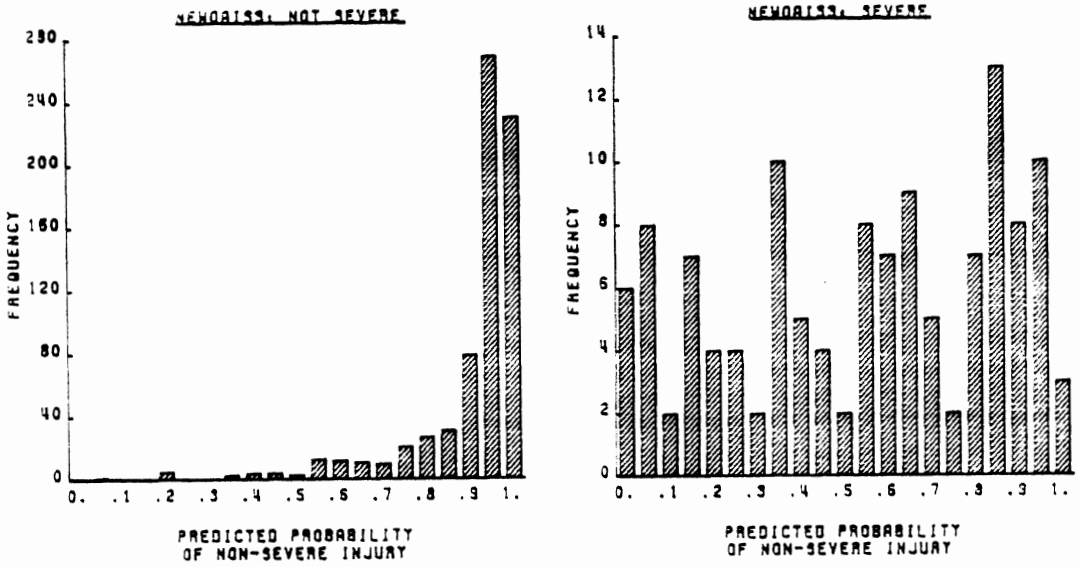

FIGURE 3.71 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For CIA-2VEH Phase 1 Data - Front Impacts

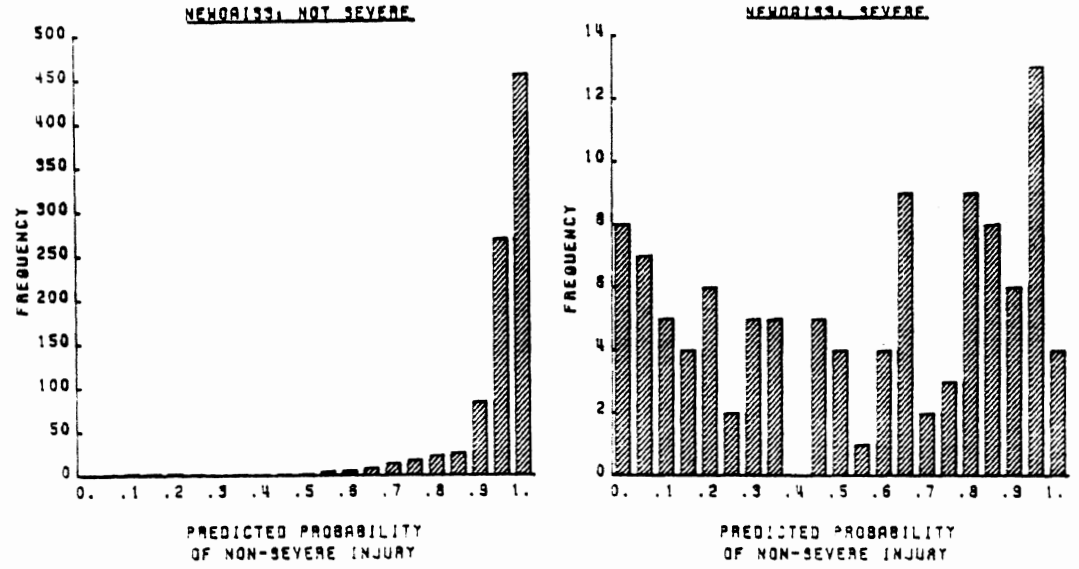

FIGURE 3.72 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OID-2VEH Phase 1 Data - Front Impacts

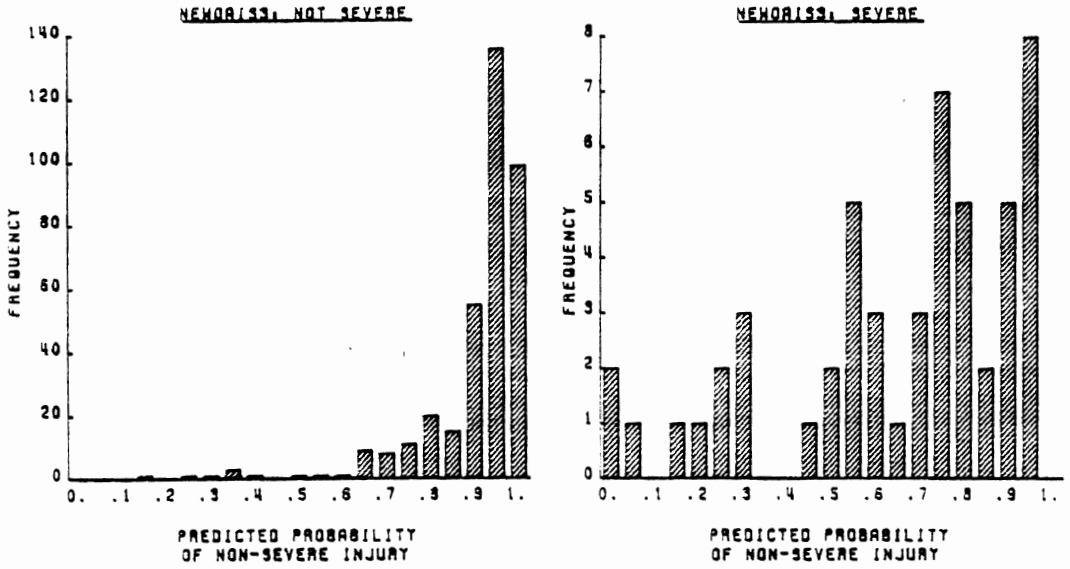

FIGURE 3.73 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OIP-2VEH Phase 1 Data - Front Impacts

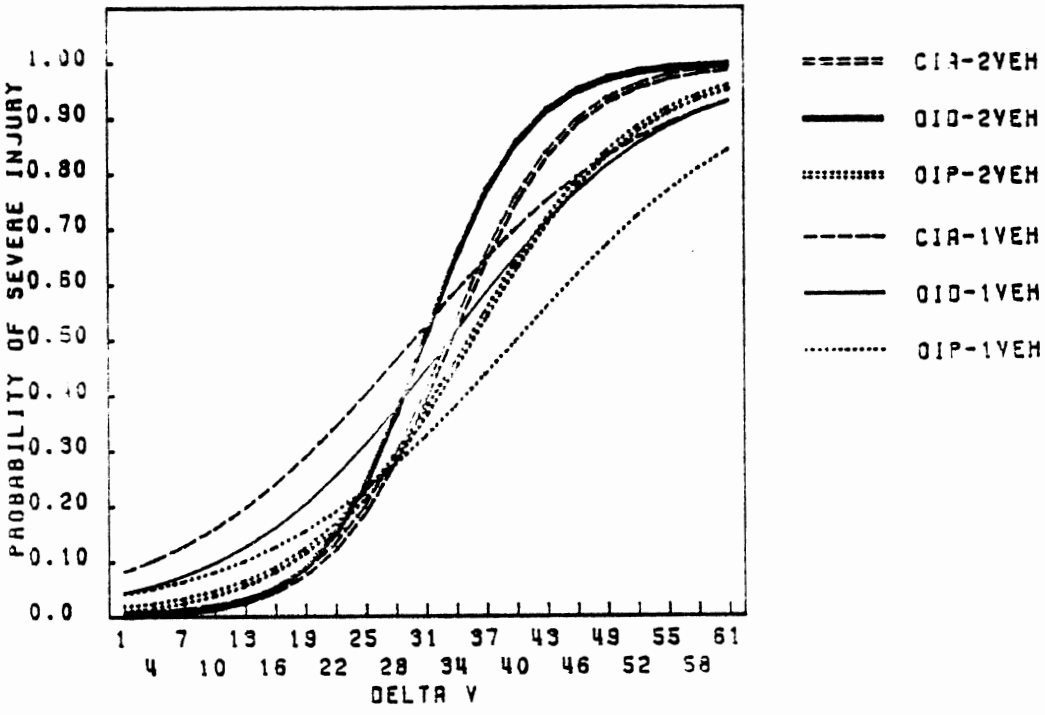

FIGURE 3.74 Logistic Curves of Two-Variable Models (Delta V, Age) For The Front-Impact Subsets Phase 1 Data - Front Impacts

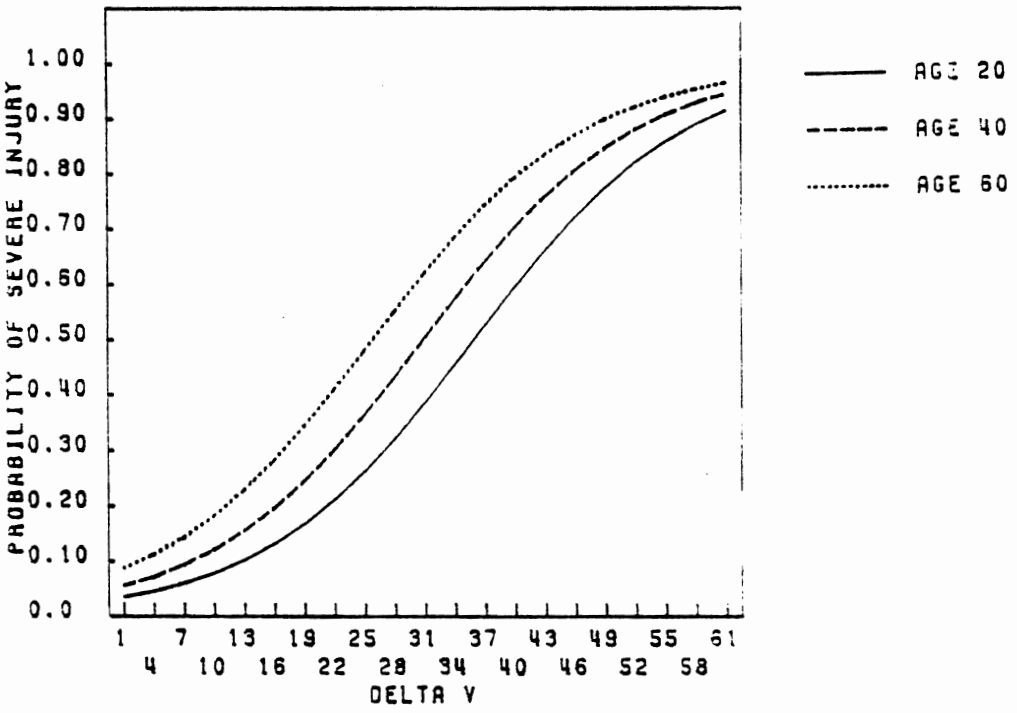

FIGURE 3.75 The Age Effect of Two-Variable Models (Delta V, Age) For OID-1VEH Phase 1 Data - Front Impacts

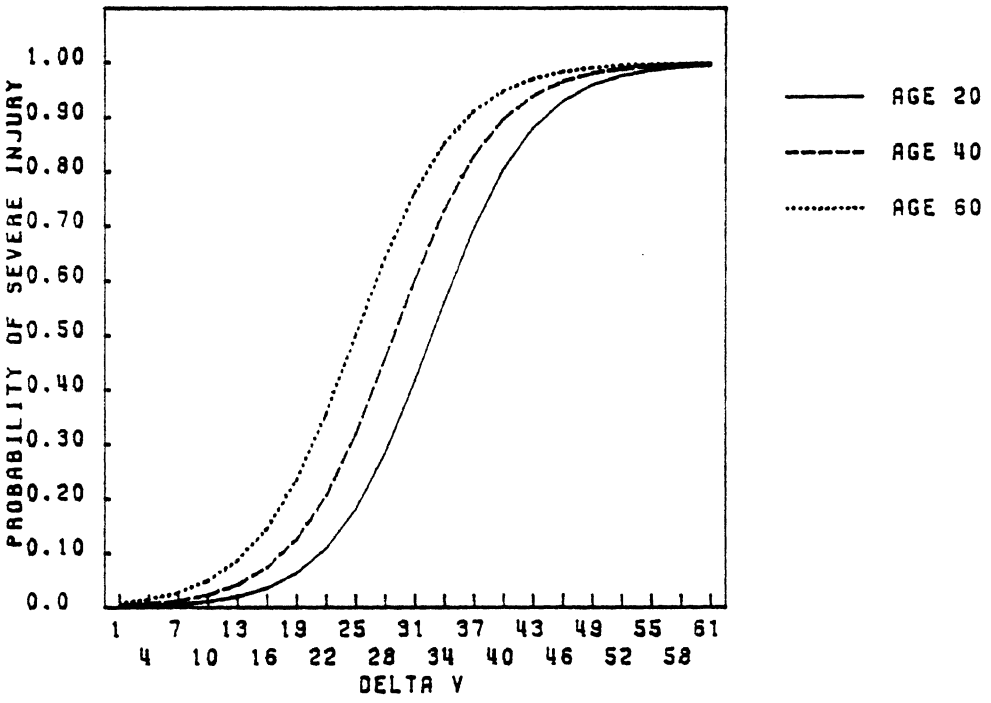

FIGURE 3.76 The Age Effect of Two-Variable Models (Delta V, Age) For OID-2VEH Phase 1 Data - Front Impacts

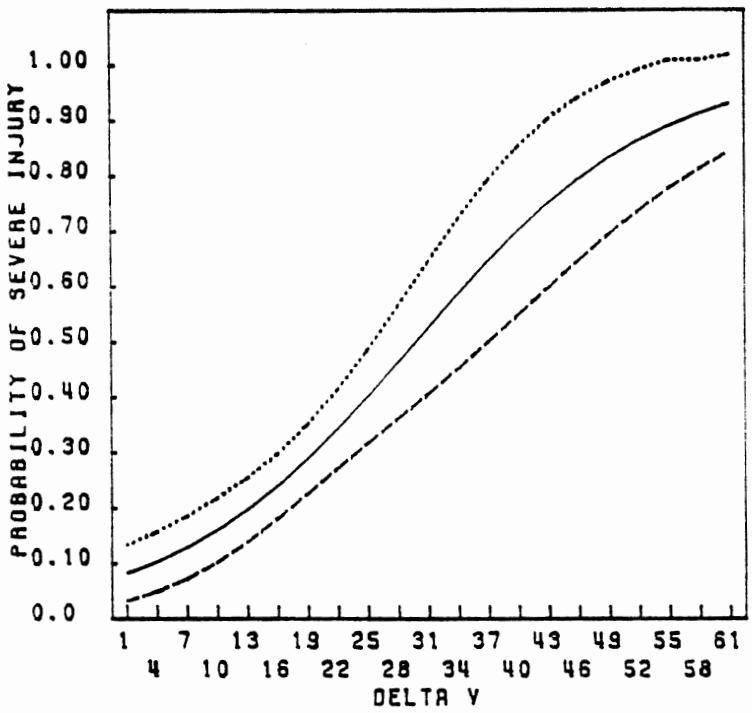

FIGURE 3.77 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For CIA-1VEH Phase 1 Data - Front Impacts

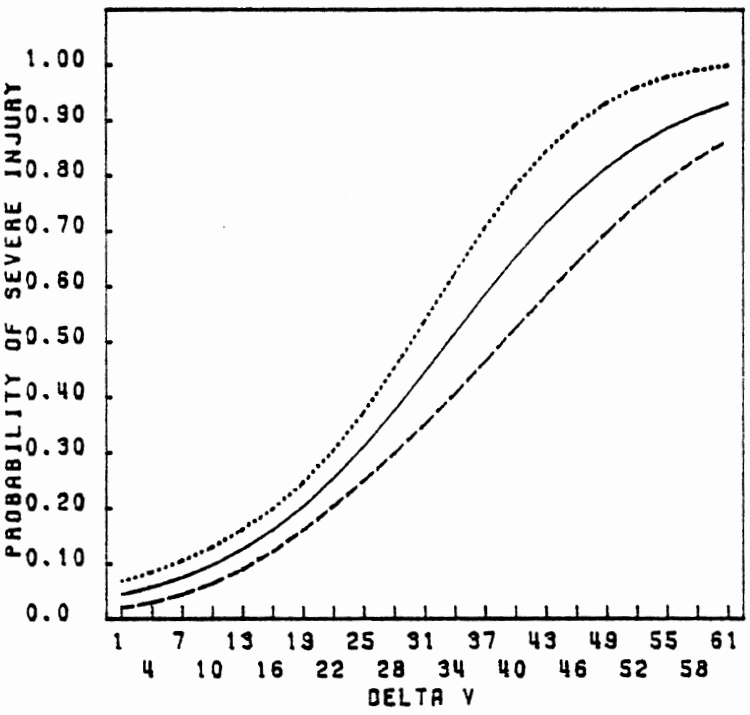

FIGURE 3.78 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For OID-1VEH Phase 1 Data - Front Impacts

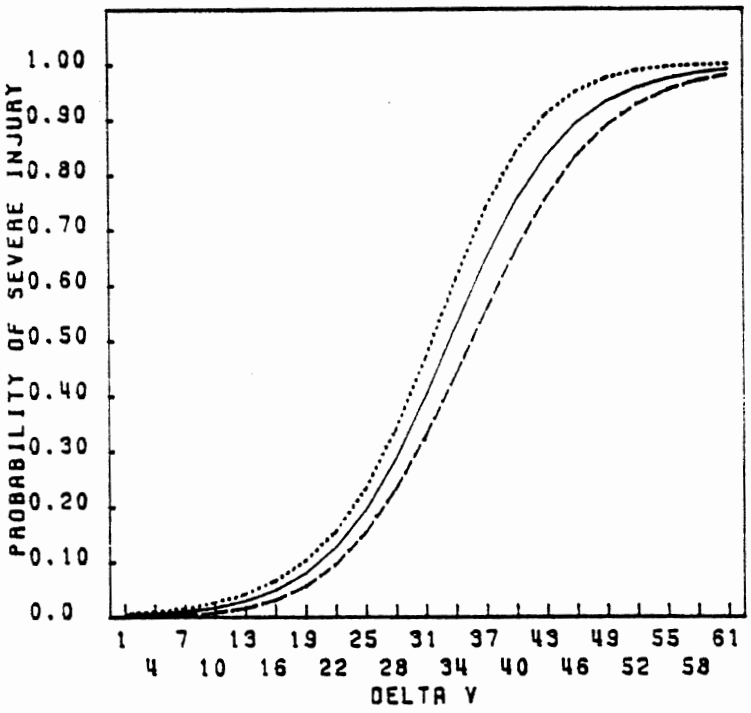

FIGURE 3.79 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For CIA-2VEH Phase 1 Data - Front Impacts

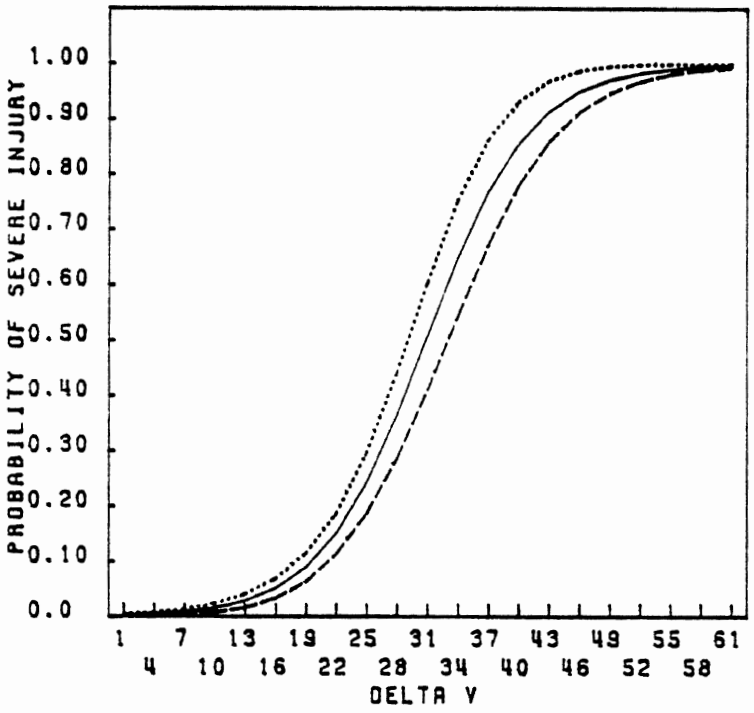

FIGURE 3.80 Confidence Interval of $\bar{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For OID-2VEH Phase 1 Data - Front Impacts

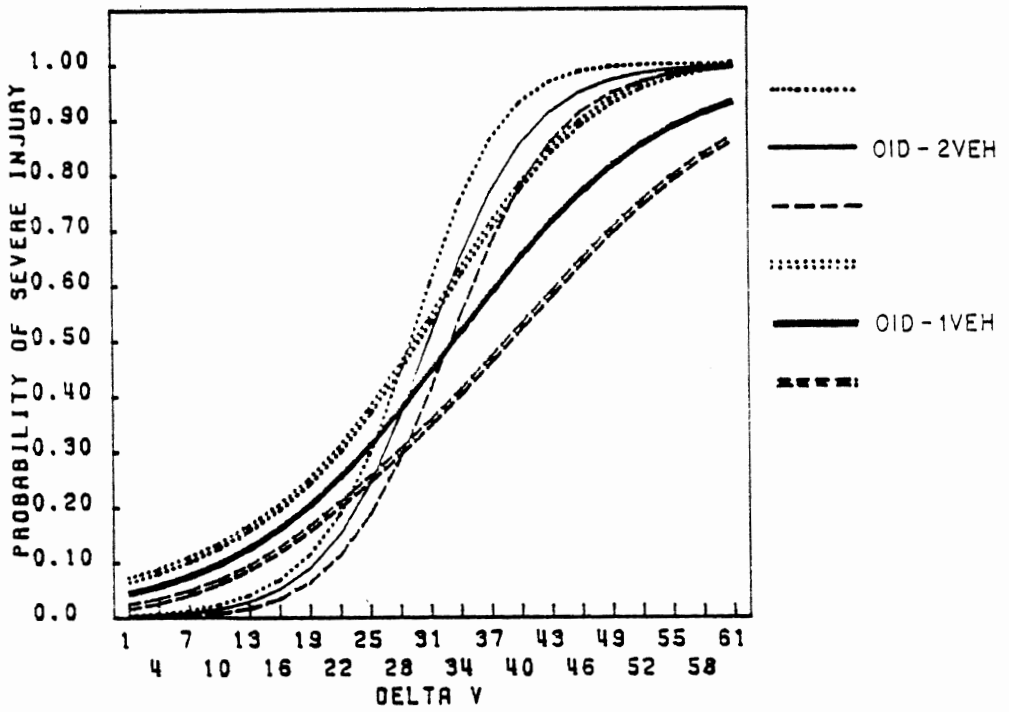

FIGURE 3.81 Confidence Intervals of $\hat{p}_i$ of Two-Variable Models (Delta V, Age) at Age 30 For OID-1VEH and OID-2VEH Phase 1 Data - Front Impacts

Estimated Models with Delta V, Age and Rural/Urban

CIA-1VEH (N=244, LRS=41.5, DF=2)

(3-74) $*\hat{p}_i = F(1.6189 - 0.0461X_1 - 0.0084X_2)$

OID-1VEH (N=434, LRS=71.1, DF=2)

(3-75) $*\hat{p}_i = F(2.1335 - 0.0522X_1 - 0.0132X_2)$

OIP-1VEH (N=189, LRS=26.6, DF=2)

(3-76) $*\hat{p}_i = F(2.4287 - 0.0443X_1 - 0.0220X_2)$

CIA-2VEH (N=849, LRS=257.0, DF=3)

(3-77) $\hat{p}_i = F(3.5721 - 0.0885X_1 - 0.0154X_2 - 0.2545X_3)$

OID-2VEH (N=1033, LRS=290.3, DF=3)

(3-78) $\hat{p}_i = F(4.0248 - 0.1037X_1 - 0.0214X_2 - 0.3837X_3)$

OIP-2VEH (N=415, LRS=87.8, DF=3)

(3-79) $\hat{p}_i = F(3.1127 - 0.0643X_1 - 0.0228X_2 - 0.3512X_3)$

where

$\hat{p}_i$ is the estimated probability of a severe injury,
F is the logistic distribution,
X_1 is Delta V,
X_2 is Age, and
X_3 is 1 if Rural and zero otherwise and
LRS is the Likelihood Ratio Statistic.

*The Rural/Urban variable was not significant.

The goodness of fit of these models is shown in Table 3.41. Rural/Urban marginally improved the percent correct prediction of severe injuries of the models with Delta V and Age.

Vehicle Weight, when incorporated into the model in the presence of Delta V, was not statistically significant. Neither did it improve the predictive capability of the existing model.

Object Contacted was incorporated into the model in the form of a set of dummy variables. The variable did not appear to significantly improve the model's predictive capability.

TABLE 3.41
Goodness of Fit

Single-Vehicle: Severity = F(Delta V, Age)
Two-Vehicle: Severity = F(Delta V, Age, Rural/Urban)

Phase 1 Data - Front Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
CIA-1VEH*	171	73	75.8	94.2	32.9
OID-1VEH*	342	92	80.9	96.5	22.8
OIP-1VEH*	161	28	85.7	98.1	14.3
CIA-2VEH	723	126	89.2	97.6	44.4
OID-2VEH	923	110	92.8	98.6	44.5
OIP-2VEH	363	52	89.2	98.1	26.9

*The rural/urban variable was not significant.

Damage distribution was brought into the model in the form of dummy variables. It did not appear to be a significant variable.

Height was brought into the model but it did not appear to be a significant explanatory variable of injury severity. Occupant Weight was brought into the model but it was not a significant explanatory variable of injury severity. Sex also appeared to be an insignificant explanatory variable of injury severity.

The modelling results so far had indicated that the estimated models for single-vehicle accidents and two-vehicle accidents differed in that for the former Delta V and Age were found to be significant explanatory variables of injury severity, while, for the latter Delta V, Age and Rural/Urban were found to be significant. The estimated model had enabled us to correctly predict the low severity injuries well above 95% of the time and to correctly predict the high severity injuries at best about 45% of the time. The next step was to examine the outliers (mispredictions) by investigating variables such as Intrusion, Restraint Usage, Ejection, Injury Type and Body Region with a view to bringing

about ways and means of improving the existing model's predictive capability, particularly for the high severity injuries.

Restraint Usage was brought into the model in two ways:

1. As one variable with several levels.

2. As a dummy variable with two levels, namely, No Restraint Usage and Restraint Usage.

This variable did not improve the models' predictive capability.

The Ejection Variable was brought into the model in two ways:

1. As a dummy variable with three levels - no ejection, trapped, and others.

2. As a two-level dummy variable - trapped and others.

No significant improvement in the model predictive capability was noted although the two-level dummy, as indicated by the likelihood ratio statistic (LRS), appeared to be statistically significant.

Intrusion. Intrusion was brought into the model in two ways:

1. as a three-level variable, i.e., no intrusion; intrusion including steering column, roof,and combination of intrusions; and the rest. Such regrouping of the levels of the intrusion variable was based on combining the levels to form one class for no intrusion, one class for cases with intrusion which poorly predicted by existing models, and one class for cases with intrusion which were correctly predicted by the existing models.

2. as a two-level dummy variable, i.e., no intrusion had a value of 1 and intrusion had a value of 0.

Both methods indicated that intrusion appeared to be a significant explanatory variable of injury severity. The improvement in the model predictive capability was comparable for both methods: small. As a result, the two-level dummy variable form was chosen because of its simple form.

3.4.4 Final Models The results of the model estimation having Delta V, Age, Rural/Urban and the two-level Intrusion as the independent variables are shown below:

Estimated Models with Delta V, Age, Rural/Urban, No Intrusion/Intrusion

CIA-1VEH (N=241, LRS=59.5, DF=3)

(3-80) $\hat{p}_i = F(0.8501 - 0.0284X_1 - 0.0090X_2 + 0.8274X_4)$

OID-1VEH (N=434, LRS=101.6, DF=3)

(3-81) $\hat{p}_i = F(1.4601 - 0.0329X_1 - 0.0149X_2 + 0.9772X_4)$

OIP-1VEH (N=137, LRS=33.4, DF=3)

(3-82) $\hat{p}_i = F(1.9959 - 0.0321X_1 - 0.0253X_2 + 0.8392X_4)$

CIA-2VEH (N=841, LRS=277.7, DF=4)

(3-83) $\hat{p}_i = F(2.8134 - 0.0723X_1 - 0.0170X_2 - 0.1993X_3 + 0.7133X_4)$

OID-2VEH (N=1020, LRS=313.0, DF=4)

(3-84) $\hat{p}_i = F(3.1038 - 0.0811X_1 - 0.0225X_2 - 0.2478X_3 + 0.8600X_4)$

OIP-2VEH (N=412, LRS=92.8, DF=4)

(3-85) $\hat{p}_i = F(2.5358 - 0.0501X_1 - 0.0236X_2 - 0.2704X_3 + 0.5613X_4)$

where

$\hat{p}_i$ is the estimated probability of a severe injury,
F is the logistic distribution,
X_1 is Delta V,
X_2 is Age,
X_3 is 1 if Rural zero otherwise, and
X_4 is 1 if no intrusion and zero otherwise, and
LRS is the Likelihood Ratio Statistic.

The goodness of fit of these models is shown in Table 3.42. By having Intrusion in the model, the predictive capability of the models, particularly in predicting severe injuries, improved considerably.

The estimated logistic curves are shown in Figure 3.82. These curves show how the probability of a severe injury varies with Delta V values holding Age fixed at 30 and no intrusion was involved. Notice the difference between this figure and Figure 3.74 which are based on the models with Delta V, Age, No Intrusion/Intrusion and Rural/Urban and the models with Delta V and Age respectively. The "No Intrusion" is such that it lowers the probability of a severe injury for both single-

TABLE 3.42
Goodness of Fit
Single-Vehicle: Severity = F(Delta V, Age, Intrusion)
Two-Vehicle: Severity = F(Delta V, Age, Rural/Urban, Intrusion)

Phase 1 Data - Front Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
CIA-1VEH	169	72	78.8	91.1	50.0
OID-1VEH	342	92	81.6	95.0	31.5
OIP-1VEH	159	28	87.2	98.7	21.4
CIA-2VEH	716	125	90.2	96.8	52.8
OID-2VEH	910	110	93.0	98.1	50.9
OIP-2VEH	360	52	89.3	97.2	34.6

vehicle and two-vehicle subsets for all values of Delta V. The effect of Age is shown in Figure 3.83 for the OID-1VEH subset and in Figure 3.84 for the OID-2VEH subset. Each figure consists of three curves representing Age 20, 40 and 60. The curves show how the estimated probability of a severe injury ($1-\hat{p}_i$) varies with Delta V. For both subsets the Age effect is such that an older occupant is expected to have a higher probability of a severe injury than a younger occupant. For the two-vehicle subset, the Age effect approaches zero as Age becomes either very small or very large. This is not necessarily so with the single-vehicle subset. The Rural/Urban effect at Age 30 is illustrated in Figure 3.85 for the OID-2VEH subset. This effect is relatively small. Figures 3.86 and 3.87 show the effect of Intrusion for the OID-1VEH and OID-2VEH subsets respectively. A large intrusion effect was estimated for the single-vehicle subset, while this effect is quite small in the two-vehicle subset. Finally, confidence limits are shown as a function of Delta V for the CIA-1VEH, OID-1VEH, CIA-2VEH, and OID-2VEH subsets in Figures 3.88 - 3.91, respectively. The confidence limits are large for the single-vehicle subsets while those for the two-vehicle subsets are relatively narrow and approach zero for very small or large values of Delta V. Figure 3.92 compares the confidence limits

of the OID-1VEH and the OID-2VEH models. An occupant in a two-vehicle accident is expected to have a much higher probability of a severe injury than an occupant in a single-vehicle accident when Delta V values are greater than 30 mph. The narrower confidence limits of the two-vehicle subset also implies that the prediction by the two-vehicle model is likely to be more reliable than that by the single-vehicle model.

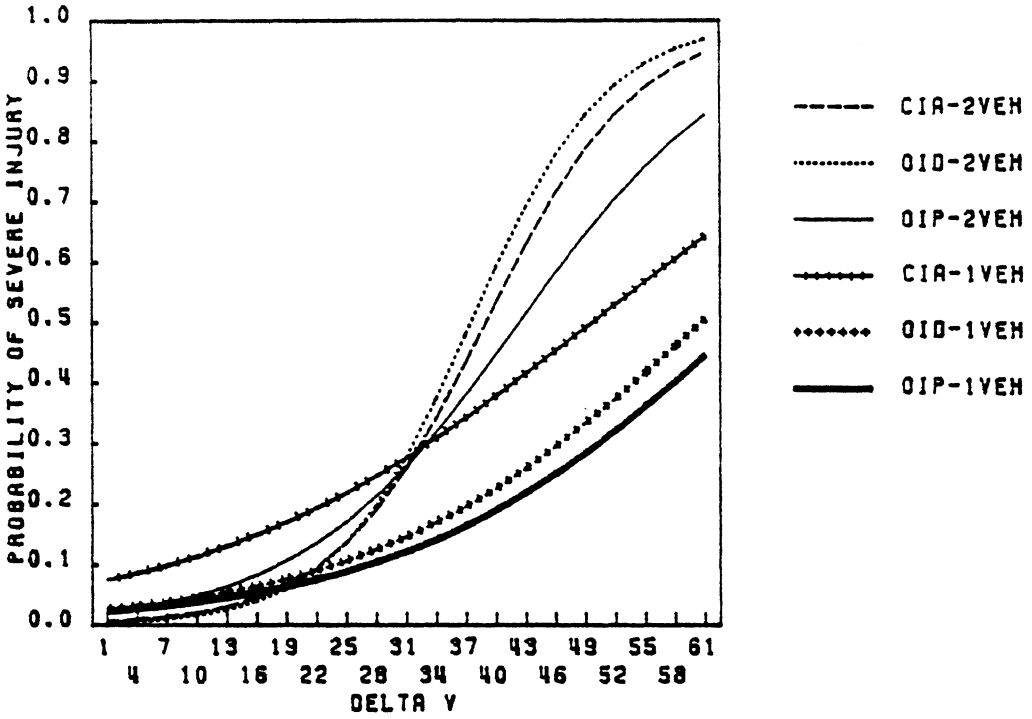

FIGURE 3.82 Logistic Curves of Four-Variable Models (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For Front-Impact Subsets Phase 1 Data - Front Impacts

Interaction Effects. The modelling results could be summarized as follows:

Single-vehicle

Severity = F(Delta V, Age, Dummy [No Intrusion/Intrusion])

Two-vehicle

Severity = F(Delta V, Age, Dummy [Rural/Urban], Dummy [No Intrusion/Intrusion])

The presence of dummy variables in the above mentioned forms changed the constant terms of the estimated models but not the estimated

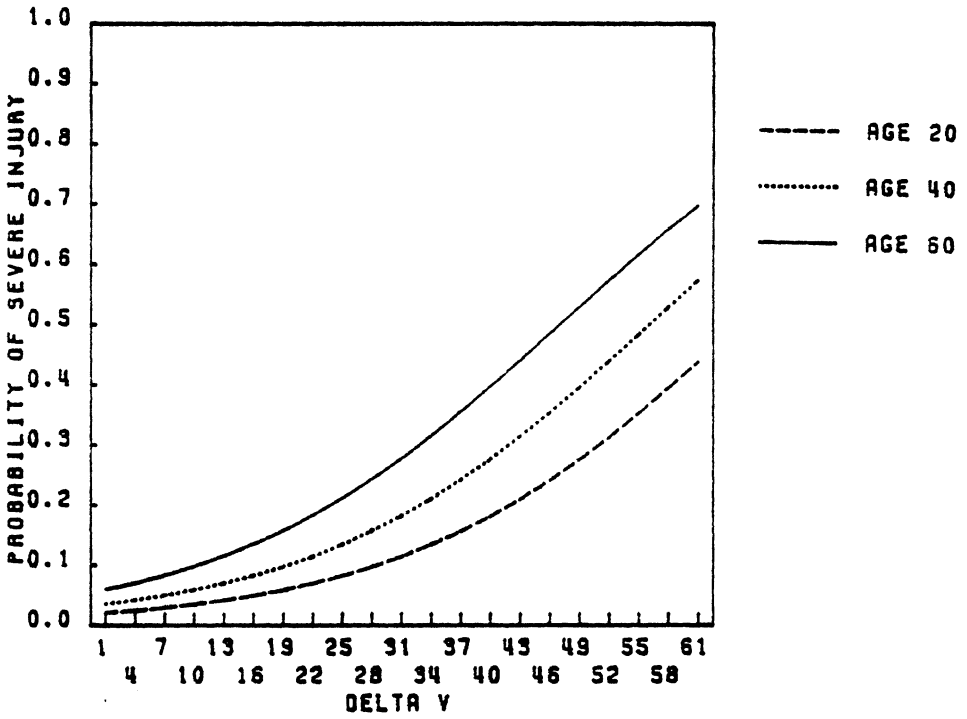

FIGURE 3.83 The Age Effect of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For OID-1VEH Phase 1 Data - Front Impacts

coefficients of Delta V and Age. Interaction effects of the dummy intrusion variable in the models were examined as follows.

1. Instead of the dummy variable as before, an interaction variable involving the dummy variable and Delta V was introduced into the modelling. The resultant goodness of fit did not improve although the interaction term showed statistical significance.

2. Both the dummy variable and the interaction variable were included in the modelling. The resultant goodness of fit, by and large, decreased and neither the dummy terms nor the interactive terms were statistically significant.

3. A similar procedure was repeated but this time the interaction variable was in the form of the dummy variable multiplied by CDC Extent. Similar conclusions as above were reached.

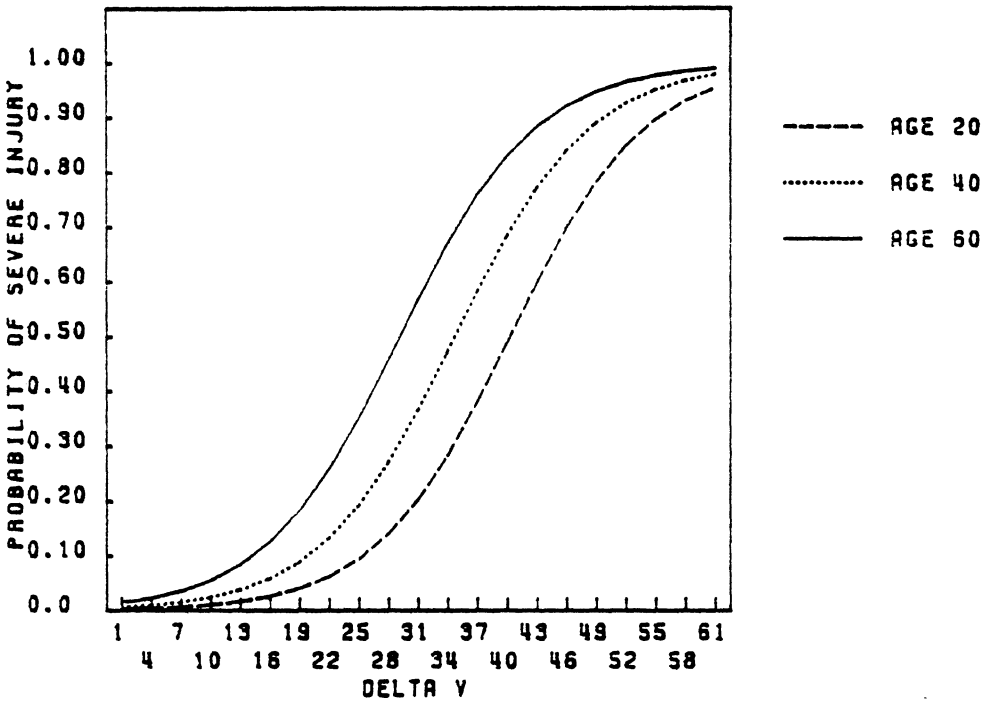

FIGURE 3.84 The Age Effect of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For OID-2VEH Phase 1 Data - Front Impacts

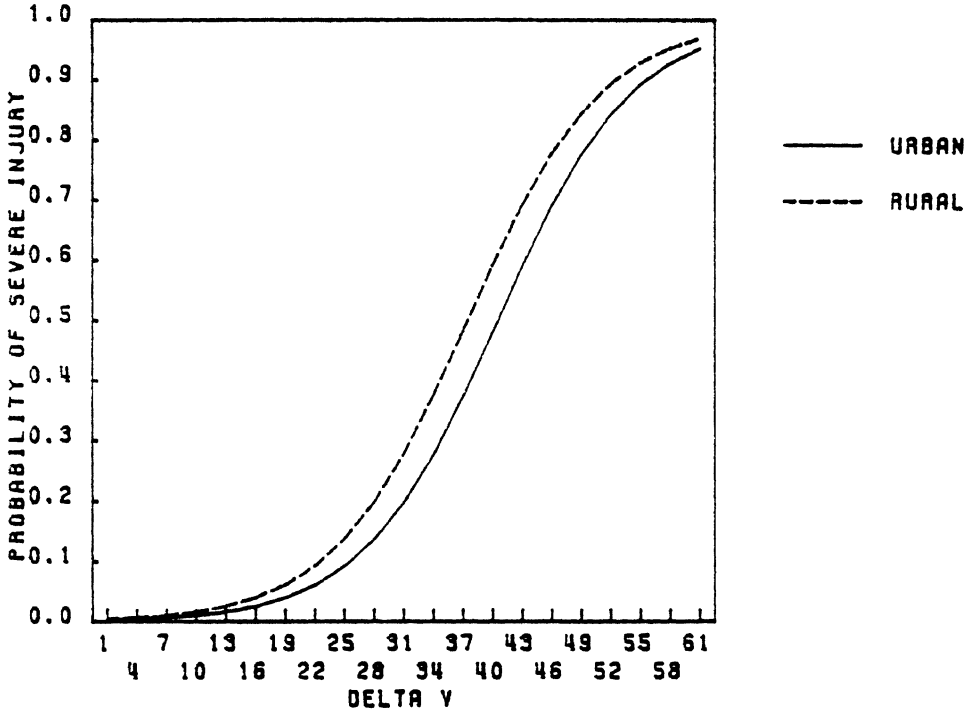

FIGURE 3.85 The Rural/Urban Effect of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For OID-2VEH Phase 1 Data - Front Impacts

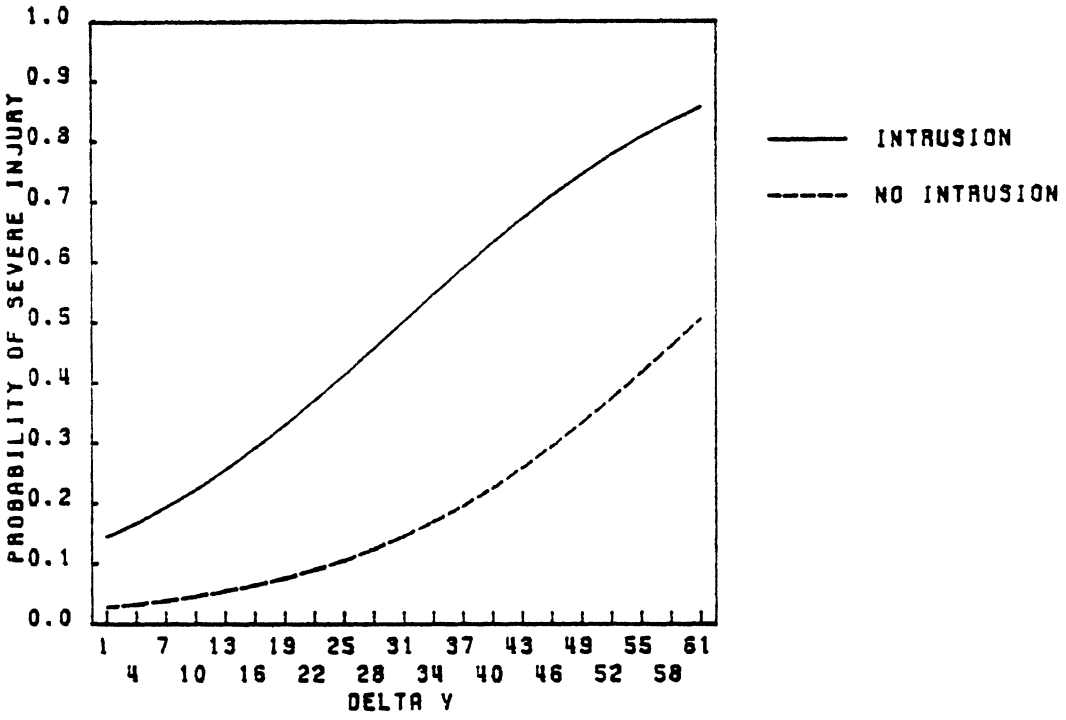

FIGURE 3.86 The Intrusion Effect of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For OID-1VEH Phase 1 Data - Front Impacts

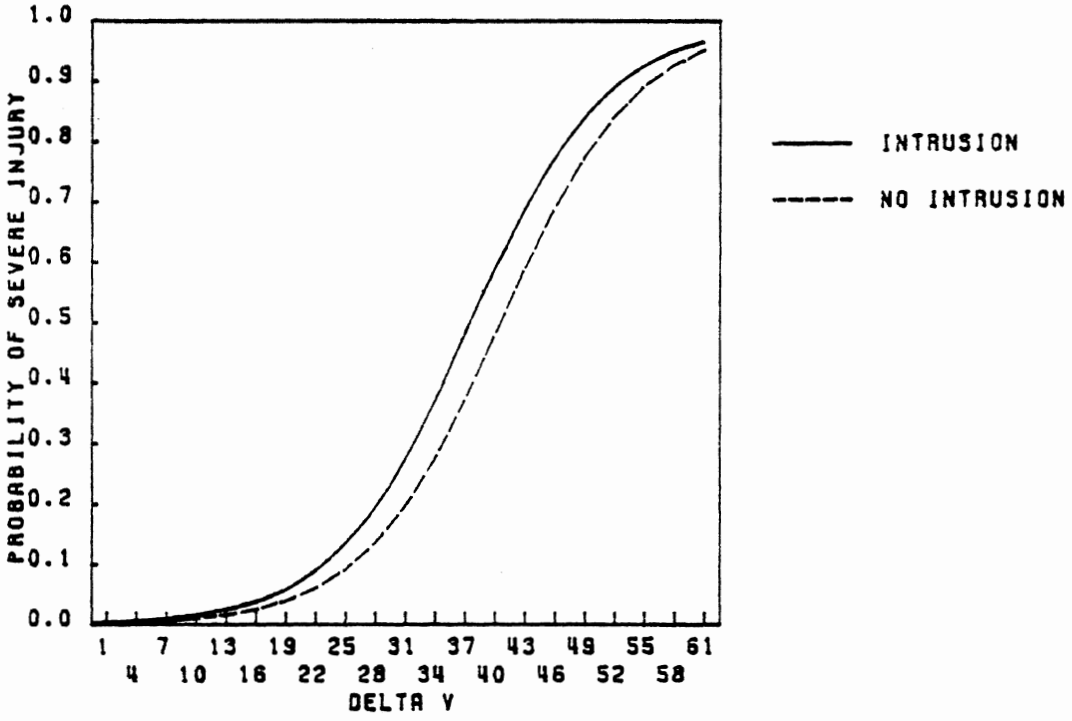

FIGURE 3.87 The Intrusion Effect of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For OID-2VEH Phase 1 Data - Front Impacts

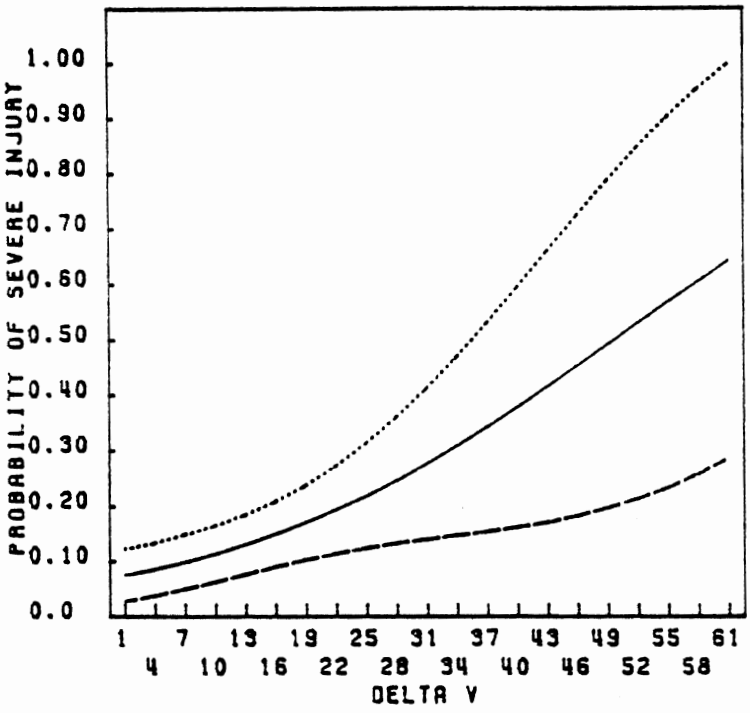

FIGURE 3.88 Confidence Interval of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) at Age 30 For CIA-1VEH Phase 1 Data - Front Impacts

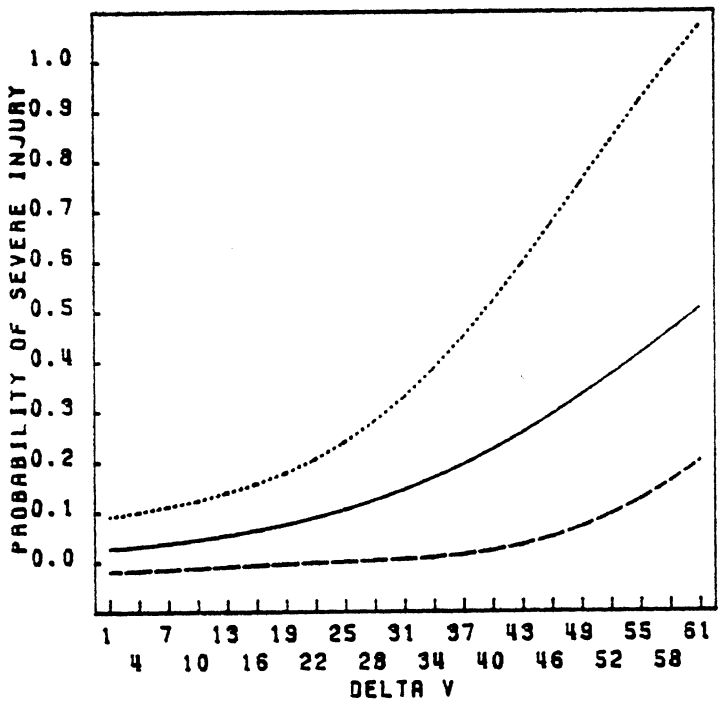

FIGURE 3.89 Confidence Interval of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) at Age 30 For OID-1VEH Phase 1 Data - Front Impacts

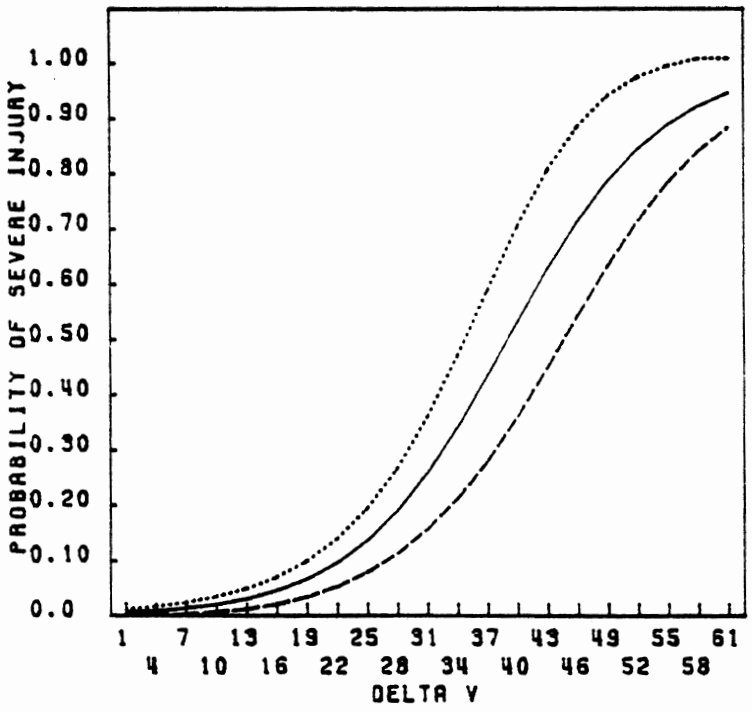

FIGURE 3.90 Confidence Interval of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) at Age 30 For CIA-2VEH Phase 1 Data - Front Impacts

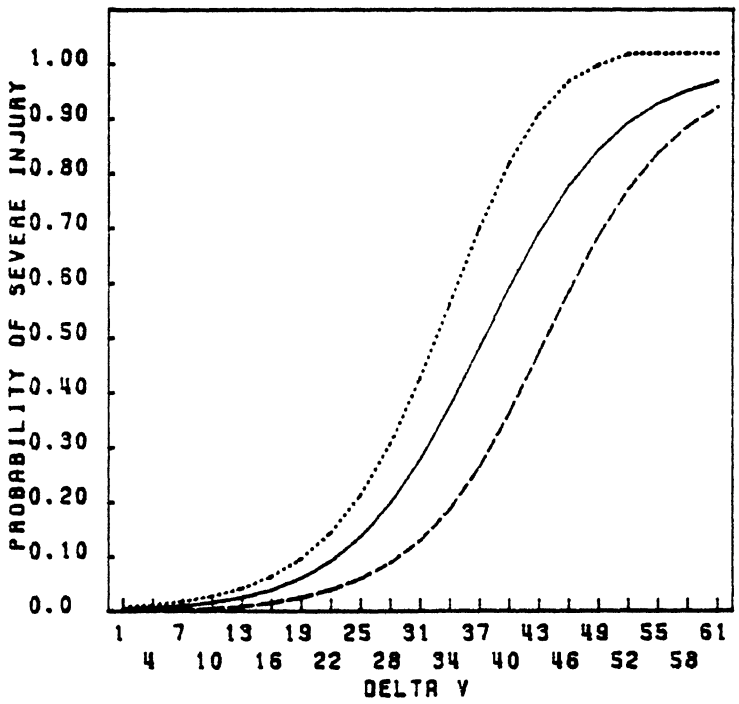

FIGURE 3.91 Confidence Interval of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) at Age 30 For OID-2VEH Phase 1 Data - Front Impacts

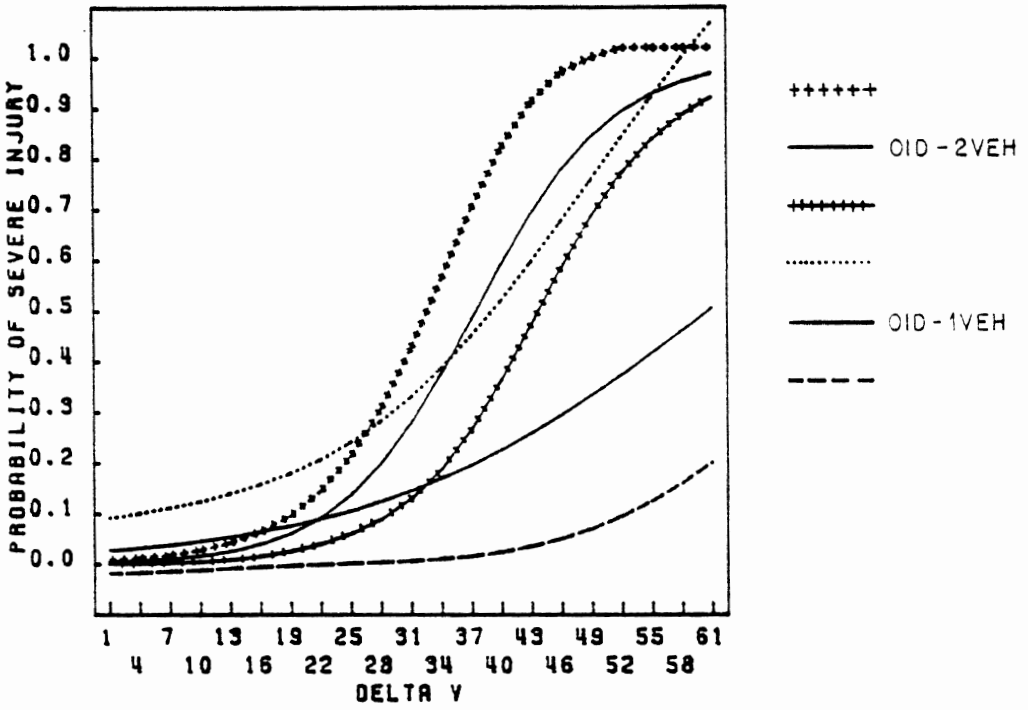

FIGURE 3.92 Confidence Interval of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) at Age 30 For OID-1VEH and OID-2VEH Phase 1 Data - Front Impacts

3.4.5 Model Evaluation. As indicated by Table 3.42, the estimated models represented by Equations 3-80 to 3-85 were capable of predicting the non-severe injuries very well but they were mispredicting the severe injuries 40% to 80% of the time. For each subset, the $\hat{p}_i$ values were calculated and two histograms of these $\hat{p}_i$ values were plotted, one for the non-severe injuries and the other for the severe injuries (Figures 3.93 to 3.98). Ideally, a good model should display, for non-severe cases, the $\hat{p}_i$ values greater than 0.5 and clustering around 0.9 to 1.0; and for severe cases, the $\hat{p}_i$ values smaller than 0.5 and close to zero. Figures 3.93 to 3.98 show that the estimated models are almost perfect for non-severe injuries but much less so for severe injuries. For the latter, about 26% of the $\hat{p}_i$ values of CIA-1VEH, OID-1VEH, CIA-2VEH and OID-2VEH and about 45% of the $\hat{p}_i$ values of OIP-1VEH and OIP-2VEH, which should have been less than 0.5, were greater than 0.75. These outliers were those cases where Delta V values were low to moderate but the resultant injuries were coded severe. This seems to imply that certain injury types could have been severe and/or certain body regions could have sustained severe injuries even though Delta V (i.e., the crash severity) was relatively low.

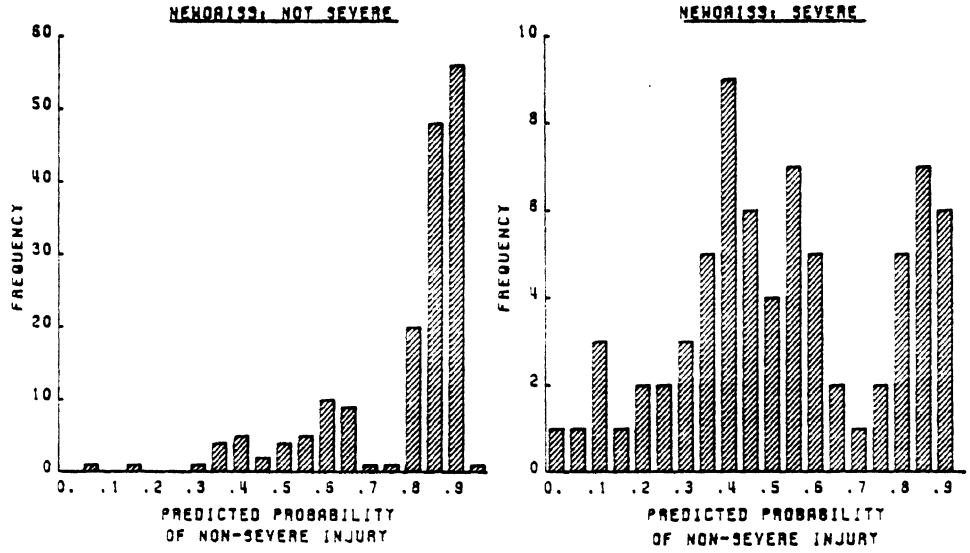

FIGURE 3.93 Histograms of $\hat{p}_i$ of The Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For CIA-1VEH Phase 1 Data - Front Impacts

Further investigation on NEWOAIS3 coding in relation to the incurred injury types and body regions revealed the following:

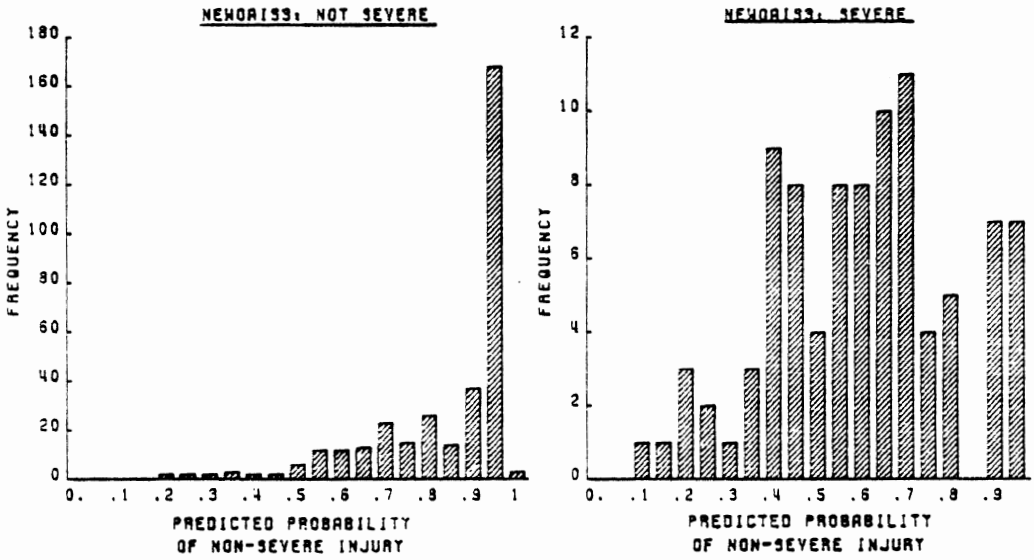

FIGURE 3.94 Histograms of $\hat{p}_i$ of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For OID-1VEH Phase 1 Data - Front Impacts

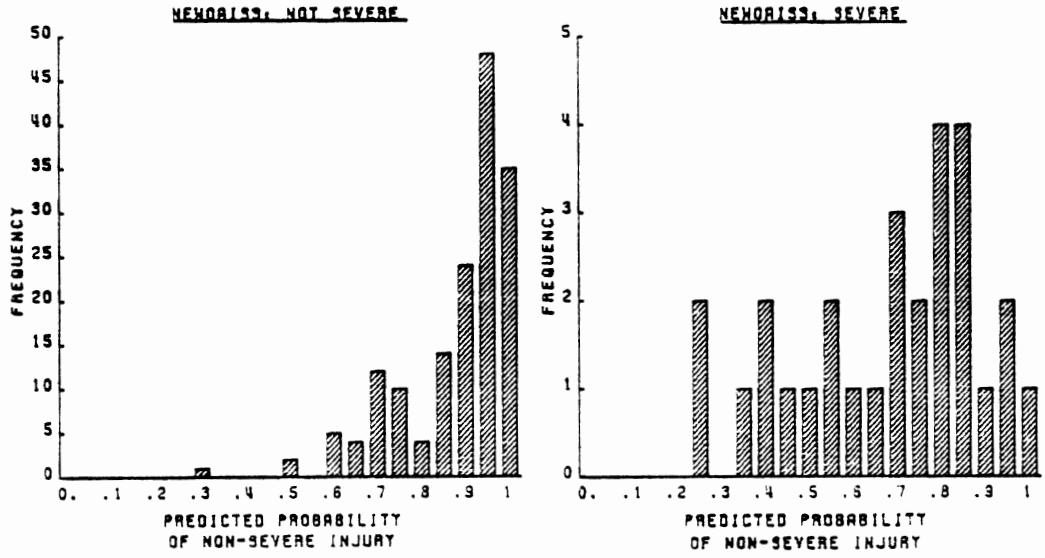

FIGURE 3.95 Histograms of $\hat{p}_i$ of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For OIP-1VEH Phase 1 Data - Front Impacts

1. The injury types which had high proportions of severe injuries to total injuries were rupture, dislocation, crushing, fracture, and hemorrhage. The first two, when they occurred, had nearly 100% severe injuries. crushing had a greater than 80% severe injuries, and for the last two about 50%.

2. The body regions which were associated with high proportions of severe injuries to total injuries were abdomen, lower extremities, chest, pelvic/hip, and knee.

On a subset-by-subset basis, the following were observed:

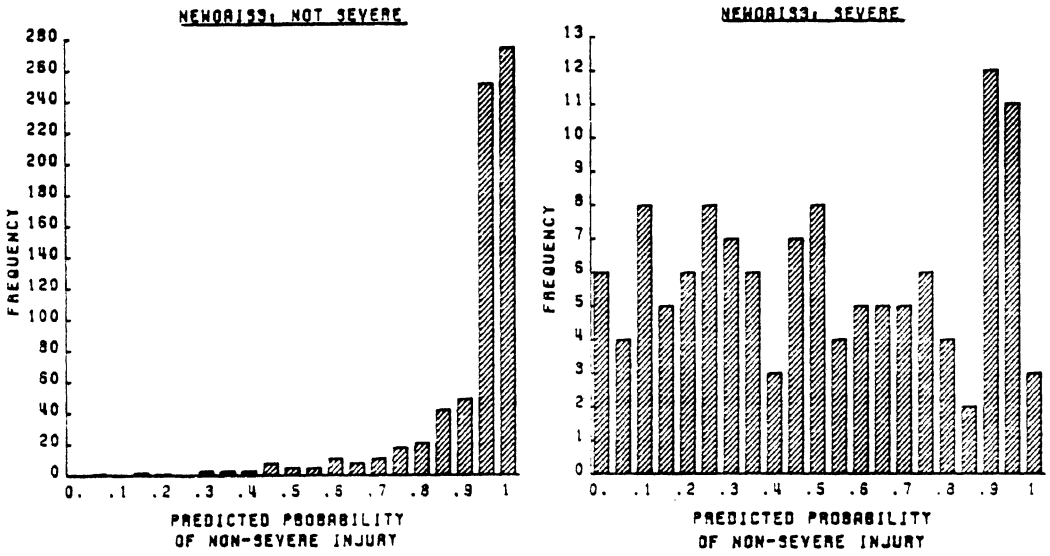

FIGURE 3.96 Histograms of $\hat{p}_i$ of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For CIA-2VEH Phase 1 Data - Front Impacts

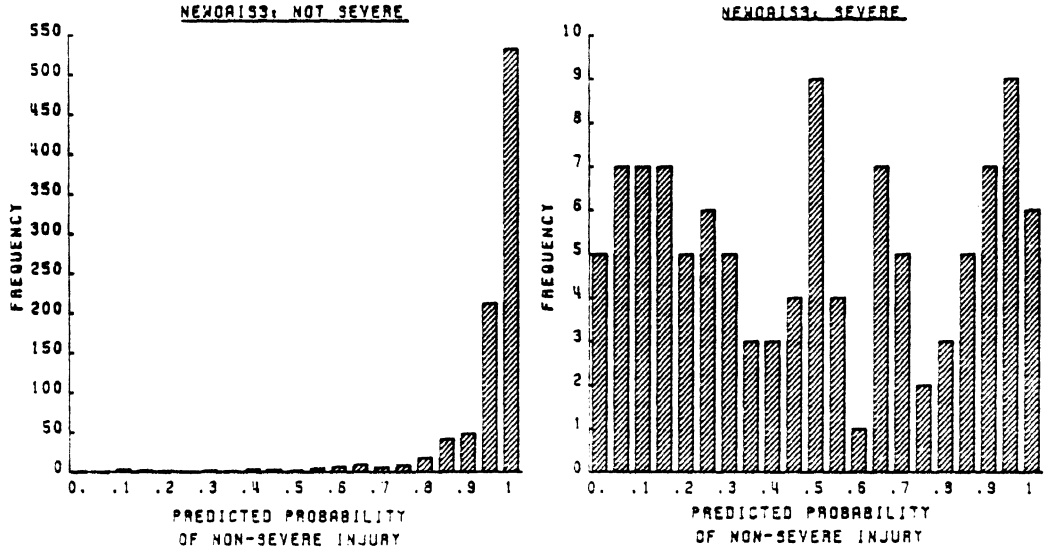

FIGURE 3.97 Histograms of $\hat{p}_i$ of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For OID-2VEH Phase 1 Data -Front Impacts

CIA-1VEH:

1. Ruptures immediately implied severe injuries and abdominal injuries only.

2. Severe dislocation injuries were more common than non-severe dislocations; the former involved pelvic/hip, ankle/foot and elbow while the latter involved wrist/hand.

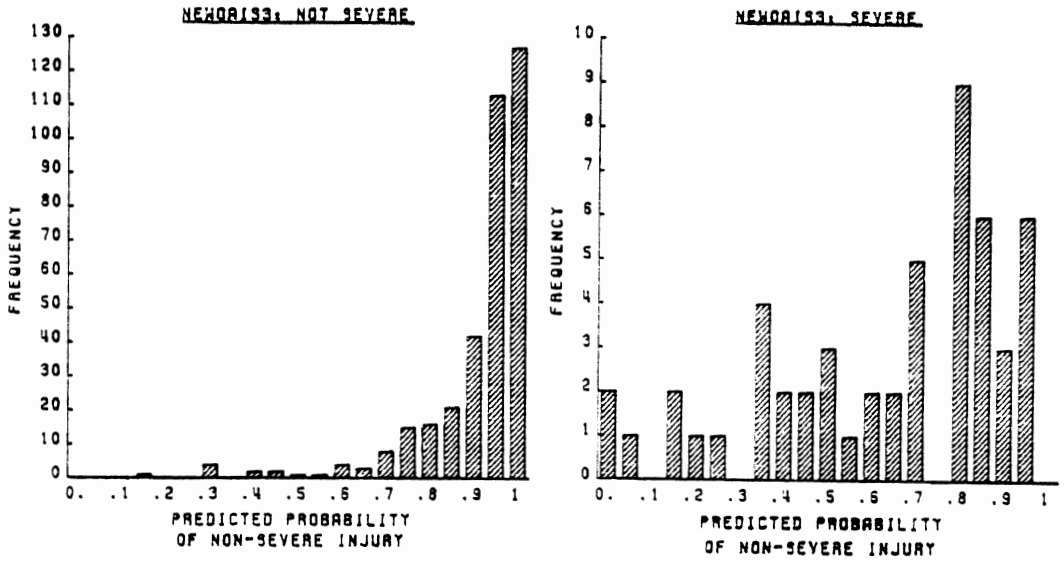

FIGURE 3.98 Histograms of $\hat{p}_i$ of Four-Variable Model (Delta V, Age, Rural/Urban and No Intrusion/Intrusion) For OIP-2VEH Phase 1 Data - Front Impacts

3. Fractures had a 50-50 chance of severe injuries.

4. Injuries involving abdomens, lower extremities immediately implied severe injuries.

OID-1VEH:

1. Ruptures and dislocations immediately implied severe injuries; the former involved only abdomens while the latter involved pelvic/hip, ankle/foot and wrist/hand.

2. Fracture had a 50-50 chance of severe injuries.

OIP-1VEH:

1. Ruptures immediately implied abdominal severe injuries.

2. Severe dislocations were far more common than non-severe dislocations, the former involved pelvic/hip and elbow while the latter involved ankle/foot.

3. Severe fractures were less common than non-severe fractures.

4. Abdomen and Lower extremities immediately implied severe injuries.

CIA-2VEH:

1. Ruptures, dislocations and hemorrhages immediately implied severe injuries.
2. Fractures had a 50-50 chance of being severe injuries.
3. Abdominal injuries were highly likely to be severe.

OID-2VEH:

1. Ruptures and hemorrhages immediately implied severe injuries. Dislocations were highly likely to be severe.
2. Fractures had a slightly higher than 50-50 chance of being severe.
3. Abdominal injuries had about a 70% chance of being severe.

OIP-2VEH:

1. Ruptures and dislocations immediately implied severe injuries.
2. Fractures and hemorrhages had a 50-50 chance of being severe.
3. Chest were more likely to be severe injuries than those to other body types.

It was also noted that hemorrhages appeared to have occurred much more frequently in two-vehicle accidents than in single-vehicle accidents.

The two-way tables of outliers and injury type and of outliers and body region were examined. Listed here are those injury types and body regions in which the proportions of outliers to total cases were high. Of particular interest from the point of view of trying to improve the prediction of the outliers were the injury types and the body regions which were ranked consistently high within the six subsets in Table 3.43. Such injury types appeared to be Dislocation, Rupture and Fracture, and to the lesser extent, Hemorrhage and Crushing, and the body regions were Lower Extremities, Abdomen, Chest, Pelvic/Hip and Forearm.

The development of mechanistic models for front impacts of the Phase 1 data had resulted in, for single-vehicle accidents, the models having Delta V, Age and No Intrusion/Intrusion as the independent variables and for two-vehicle accidents the models with Delta V, Age, No Intrusion/Intrusion and Rural/Urban. Further analyses of front impacts are continued with the addition of the Phase 2 data in Section 3.5.

TABLE 3.43
Injury Types And Body Regions with High Percentage of Misprediction

Phase 1 Data - Front Impacts

Subset	Injury Types Which Were Highly Mispredicted*	Body Regions Which Were Highly Mispredicted*
CIA-1VEH	Dislocation	Lower Extremities
	Rupture	Abdomen
	Sprain	Pelvic/Hip
	Crushing	Lower Leg/Thigh
	Avulsion	Chest
	Fracture	Neck
OID-1VEH	Dislocation	Lower Leg
	Rupture	Pelvic/Hip
	Fracture	Abdomen
		Forearm
		Chest
		Ankle/Foot
OIP-1VEH	Rupture	Abdomen
	Dislocation	Lower Extremities
	Avulsion	Upper Extremities
	Fracture	Back
		Pelvic/Hip
CIA-2VEH	Hemorrhage	Upper Extremities
	Dislocation	Lower Extremities
	Rupture	Abdomen
	Fracture	Chest
		Pelvic/Hip
		Thigh
OID-2VEH	Hemorrhage	Lower Extremities
	Rupture	Abdomen
	Dislocation	Pelvic/Hip
	Fracture	Ankle/Foot
OIP-2VEH	Rupture	Chest
	Hemorrhage	Abdomen
	Dislocation	Forearm/Arm
	Fracture	Back/Shoulder
		Pelvic/Hip
		Thigh
		Ankle/Foot

*Injury Types and body regions were ranked within subsets by the larger magnitude of misprediction proportions.

3.5 Final Analytical Results for Front Impacts

The development of mechanistic models for front impacts was continued with the Phase 2 data. Again, the first consideration was the validation of the Phase 1 models with the Phase 2 data. As with the side impacts, the initial subsections deal with the combination of the Phase 1 and Phase 2 data and with the combination of subsets. Several variables are reviewed for addition to the models using the combined data, including contact point. The final models incorporate body region as in the side impact models.

3.5.1 Validation of Phase 1 Models. The estimated models as represented by Equations 3.80 to 3.85 were applied to the Phase 2 data[22], and the following goodness of fit measures for the six subsets were obtained and are shown in Table 3.44.

Comparison of Table 3.42 and Table 3.44 revealed when the Phase 1 estimated models (Delta V, Age, No Intrusion/Intrusion, Rural/Urban) were applied to the Phase 2 data file:

1. For all subsets the overall proportion of cases correctly predicted by the Phase 1 models for the Phase 2 data was only 1% to 6% lower than for the Phase 1 data.

2. With the exceptions of the non-driver-only subsets (OIP-1VEH and OIP-2VEH), the models predicted severe injuries as well in the Phase 2 data as in the Phase 1 data. For the non-driver-only subsets, the models predicted severe injuries only half as well in the Phase 2 data.

Because Delta V and to the lesser extent Age had been known to generally exert strong influence on injury severity, it was considered worthwhile to attempt to fit the Phase 2 data with the Phase 1 models that only had Delta V and Age as the independent variables and to compare the goodness of fit measures with those obtained from models

---

[22]The Phase 2 data file gives a slightly different version of intrusion information from that contained in the Phase 1 data. This had an effect in the creation of the intrusion dummy variable, that was used in determining the results contained in Table 3.44. The variable created from the Phase 2 data has a numerical value of zero if an intrusion was specified and a value of one otherwise.

TABLE 3.44

Goodness of Fit

Single Vehicle - Severity = F(Delta V, Age, Intrusion)
Two Vehicle - Severity = F(Delta V, Age, Intrusion, Rural/Urban)

Phase 2 Data - Front Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
CIA-1VEH	82	41	77.2	87.8	56.1
OID-1VEH	188	66	77.6	95.2	27.3
OIP-1VEH	96	23	81.5	100.0	4.3
CIA-2VEH	517	103	88.2	96.5	46.6
OID-2VEH	498	65	92.7	98.2	50.8
OIP-2VEH	194	35	87.3	99.5	20.0

described by Equations 3.68 to 3.73. Table 3.45 shows the proportion of cases correctly predicted by such models.

Comparison of Table 3.44 and Table 3.45 reveals that by bringing the two variables (Intrusion/No Intrusion and Rural/Urban) into the models, the overall goodness of fit only marginally improved. However, the additional variable(s), with the exception of the non-driver-only subsets, brought about a marked improvement in the prediction of the severe injuries. Such improvement was accompanied by a slight reduction in the proportion of non-severe injuries that had already been correctly predicted by the two-variable models.

Comparison of Table 3.40 and Table 3.45 reveals that the goodness of fit results of applying the Phase 1 models (Delta V and Age) to the Phase 2 data were almost as good as those of the Phase 1 data.

TABLE 3.45

Goodness of Fit

Severity = F(Delta V, Age)

Phase 2 Data - Front Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
CIA-1VEH	82	41	73.2	90.2	39.0
OID-1VEH	188	66	76.4	96.3	19.7
OIP-1VEH	96	23	81.5	100.0	4.3
CIA-2VEH	517	103	87.9	98.1	36.9
OID-2VEH	498	65	91.5	98.2	40.0
OIP-2VEH	194	35	86.9	99.5	17.1

3.5.2 Model Estimation - Phase 2 Data. Table 3.46 contains the number of cases with valid Delta V. The proportion of severe injuries to total injuries are also shown for all subsets.

The basic finding was that when the estimated Phase 1 models (that had Delta V, Age, Intrusion and Rural/Urban) were applied to the Phase 2 data, the models predicted the occurrence of severe and non-severe injuries in the Phase 2 data nearly as well as they did in the Phase 1 data. Such a finding was somewhat unanticipated. Rather, it had been expected that the models would provide a mediocre fit to the Phase 2 data because they had been estimated using solely the Phase 1 data.

This finding immediately suggested that a still better prediction of injury severity in the Phase 2 data might indeed be possible. One approach was to take these four variables (Delta V, Age, Intrusion and Rural/Urban) as being the most significant independent variables and to estimate new coefficients using only the Phase 2 data.

TABLE 3.46

Range of Delta V For The Six Subsets

Phase 2 Data - Front Impacts

Subset	Delta V				Percentage of Severe Injuries
	Sample Size	Range	Mean	S.D.	
CIA-1VEH	159	3-97	21.0	13.1	28.1
OID-1VEH	324	2-62	18.2	9.2	25.3
OIP-1VEH	158	3-44	17.0	8.4	23.1
CIA-2VEH	763	3-81	18.1	10.3	17.5
OID-2VEH	699	2-65	14.7	9.2	12.0
OIP-2VEH	305	2-57	15.3	8.6	14.0

Estimation of the coefficients of these independent variables using the Phase 2 data was done. The Intrusion/No Intrusion and Rural/Urban variables were found not to be significant in the presence of Delta V and Age. Neither did they significantly improve the predictive capability of the models with only Delta V and Age (Equations 3-86 to 3-91). Note that the information on intrusion was recorded differently in the Phase 1 and Phase 2 data. Such difference might have contributed to the dissimilar influences of intrusion on the injury severity, i.e., intrusion was found to be quite significant in explaining the probability of injury in the Phase 1 data but was found insignificant in the Phase 2 data. Rural/Urban was included in the Phase 1 models because it enhanced the prediction although its influence was much weaker than Intrusion and Age. The models chosen to describe front impacts of the Phase 2 data are as follows:

Estimated Models with Delta V and Age: Phase 2:

CIA-1VEH (N=123, LRS=32.13, DF=2)

(3-86) $\hat{p}_i = F(2.0731 - 0.0531X_1 - 0.0168X_2)$

OID-1VEH (N=254, LRS=47.86, DF=2)

(3-87) $\hat{p}_i = F(2.3284 - 0.0483X_1 - 0.0261X_2)$

OIP-1VEH (N=122, LRS=18.17, DF=2)

(3-88) $\hat{p}_i = F(2.0483 - 0.0693X_1)$

CIA-2VEH (N=620, LRS=217.29, DF=2)

(3-89) $\hat{p}_i = F(3.7973 - 0.0992X_1 - 0.0222X_2)$

OID-2VEH (N=563, LRS=165.81, DF=2)

(3-90) $\hat{p}_i = F(3.8990 - 0.1137X_1 - 0.0173X_2)$

OIP-2VEH (N=229, LRS=51.11, DF=2)

(3-91) $\hat{p}_i = F(3.0408 - 0.0887X_1 - 0.0165X_2)$

where

$\hat{p}_i$ is the estimated probability of a non-severe injury,
F is the logistic function,
X_1 is Delta V,
X_2 is Age, and
LRS is the Likelihood Ratio Statistic.

Histograms of the estimated probability of a non-severe injury, $\hat{p}_i$, are presented in Figures 3.99 to 3.104. Each figure, representing a model of a particular subset, has two histograms, one for non-severe cases and the other for severe cases. The two histograms have the same axes, one represents the values of $\hat{p}_i$ at a 0.05 interval and the other the number of cases with the particular values of $\hat{p}_i$. The figures indicate that the models predicted non-severe injuries very well for all subsets but they did only half as well for severe injuries. The estimated logistic curves for the six subsets are shown on Figure 3.105. These curves show how the probability of a severe injury ($1-\hat{p}_i$) varies

with Delta V values with Age fixed at 30. As in the Phase 1 data, the main difference is between the single-vehicle and the two-vehicle subsets. The effect of Age is illustrated for the OID-1VEH and OID-2VEH subsets in Figures 3.106 and 3.107, respectively. The figures each contains three curves representing Age 20, 40 and 60. The curves show for these ages the probability of a severe injury ($1-\hat{p}_i$) as a function of Delta V. Both figures show that older occupants in general can be expected to have higher probabilities of severe injuries than the younger counterparts. The age effect is large in the single-vehicle subset. Confidence limits as a function of Delta V are shown in Figures 3.108 to 3.111 for subsets CIA-1VEH, OID-1VEH, CIA-2VEH, and OID-2VEH respectively. Each figure consists of three curves representing the upper bound, the lower bound and the estimated probability of a severe injury ($1-\hat{p}_i$) for each subset. The single-vehicle subsets have larger confidence intervals; the confidence intervals for the two-vehicle subsets are reasonably small and approach zero as Delta V values become very small or very large. Figure 3.112 compares the confidence intervals of the OID-1VEH and OID-2VEH subsets. The figure indicates that for Delta V less than 25 mph, single-vehicle accidents are likely to have higher probabilities of severe injuries than two-vehicle accidents but that for Delta V greater than 30 mph the reverse is true.

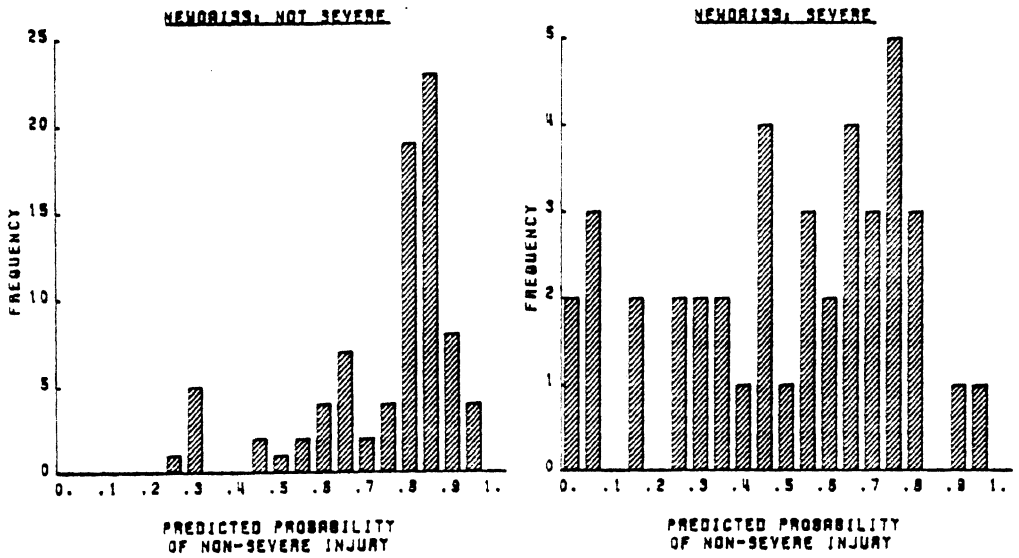

FIGURE 3.99 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For CIA-1VEH Phase 2 Data - Front Impacts

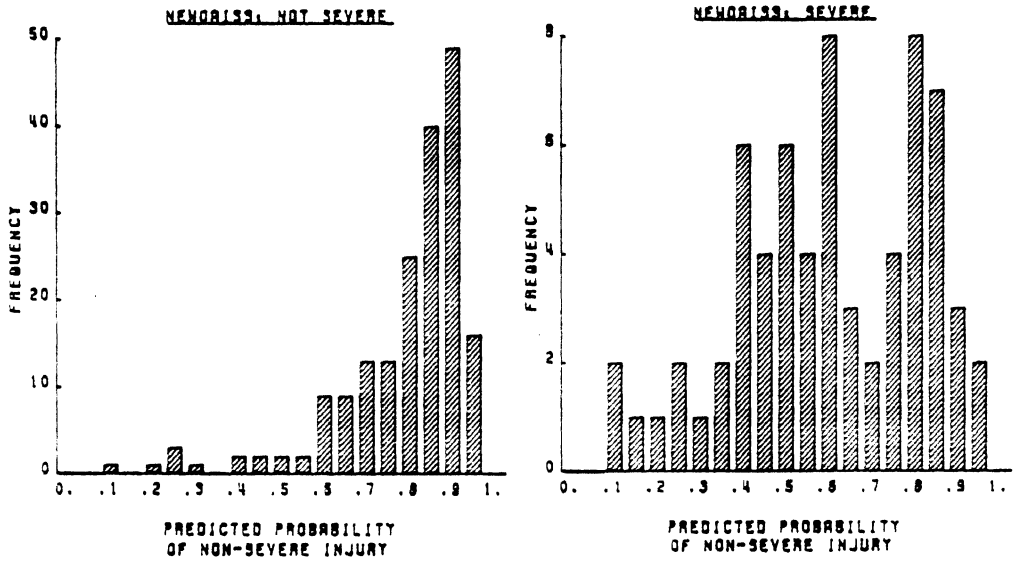

FIGURE 3.100 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OID-1VEH Phase 2 Data - Front Impacts

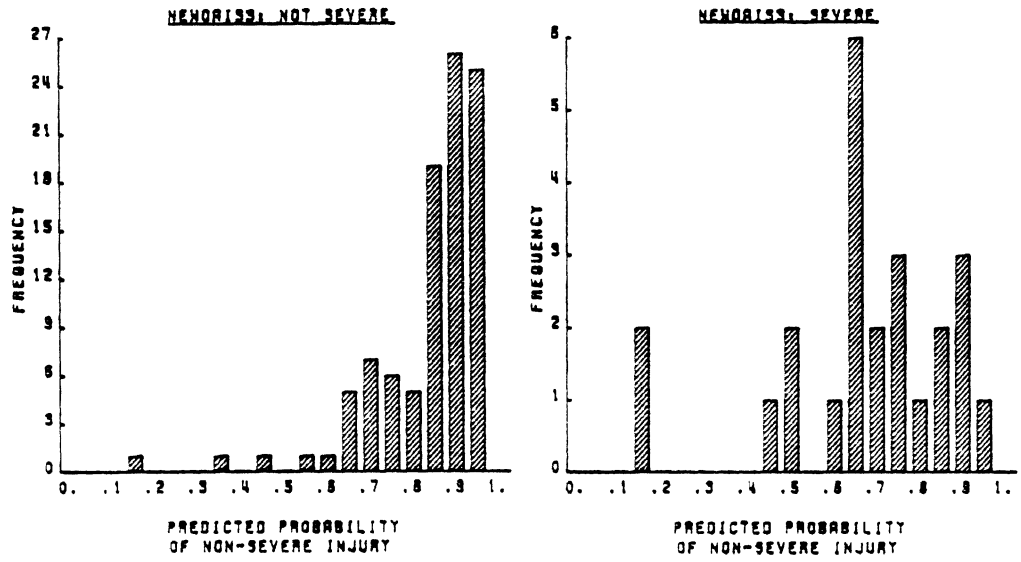

FIGURE 3.101 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OIP-1VEH Phase 2 Data - Front Impacts

The goodness of fit of the models represented by Equations 3-86 to 3-91 is contained in Table 3.47. Comparison of these goodness of fit results with those of the two-variable models (Delta V and Age) based on the Phase 1 data (Table 3.40) revealed that:

1. Their overall percentages of correct prediction were quite similar.

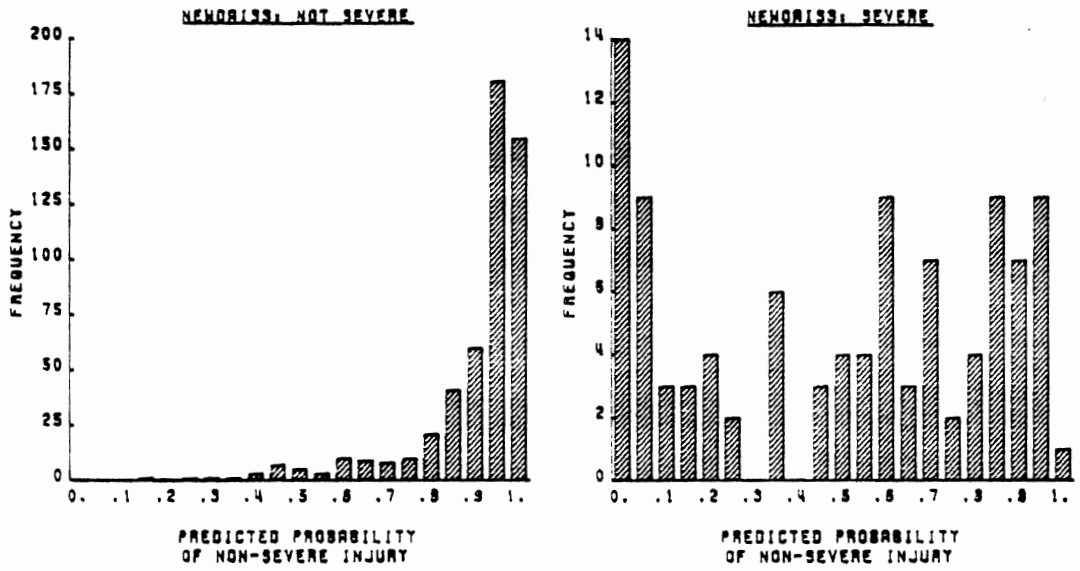

FIGURE 3.102 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For CIA-2VEH Phase 2 Data - Front Impacts

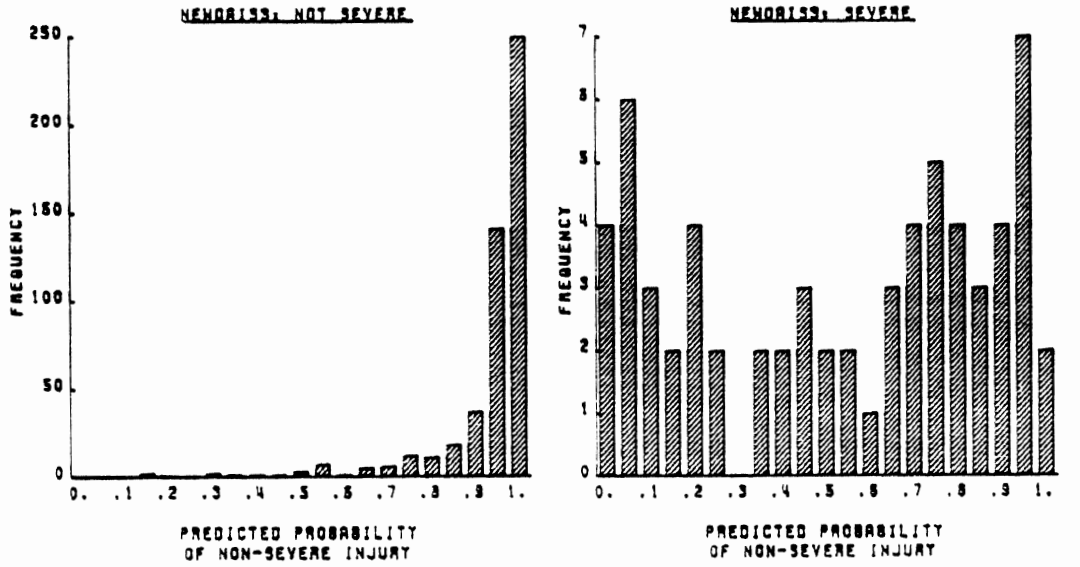

FIGURE 3.103 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OID-2VEH Phase 2 Data - Front Impacts

2. The Phase 2 data two-variable models gave higher percentages of correct prediction of severe injuries but slightly lower percentages of correct prediction of non-severe injuries than the Phase 1 data two-variable models.

Comparison of the goodness of fit results in Table 3.47 with those in Tables 3.44 and 3.45 indicate that:

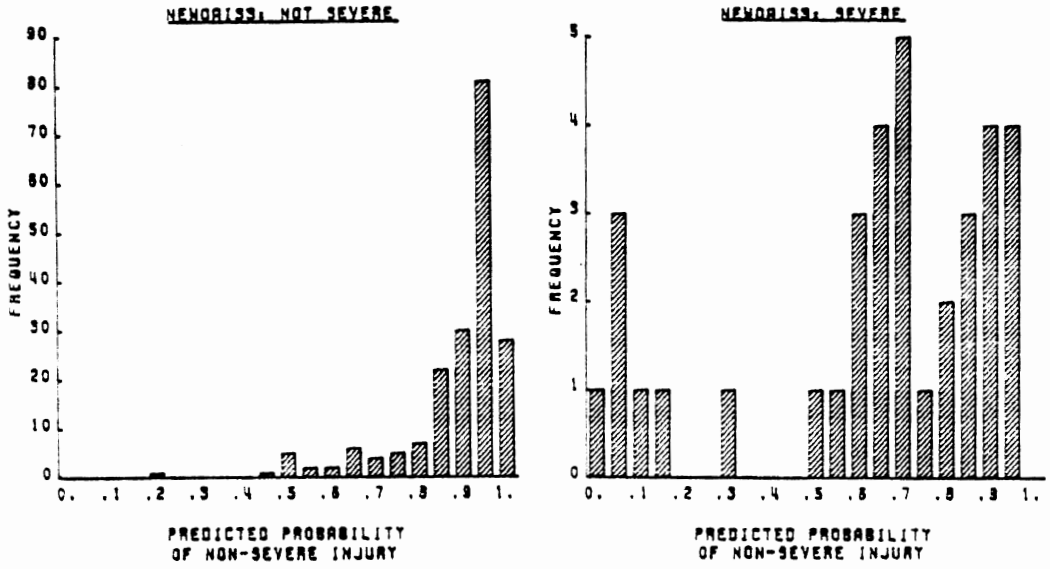

FIGURE 3.104 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OIP-2VEH Phase 2 Data - Front Impacts

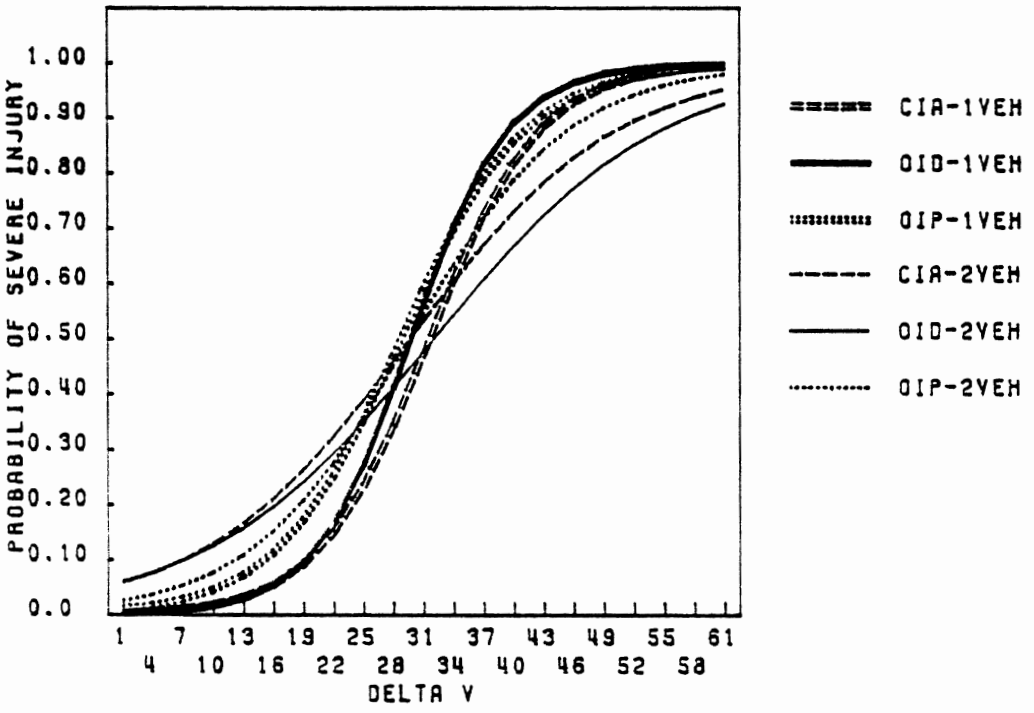

FIGURE 3.105 Logistic Curves of Two-Variable Models (Delta V, Age) For Front-Impact Subsets Phase 2 Data - Front Impacts

1. The Phase 2 data two-variable models represented by Equations 3-86 to 3-91 were predicting almost as well as when applying

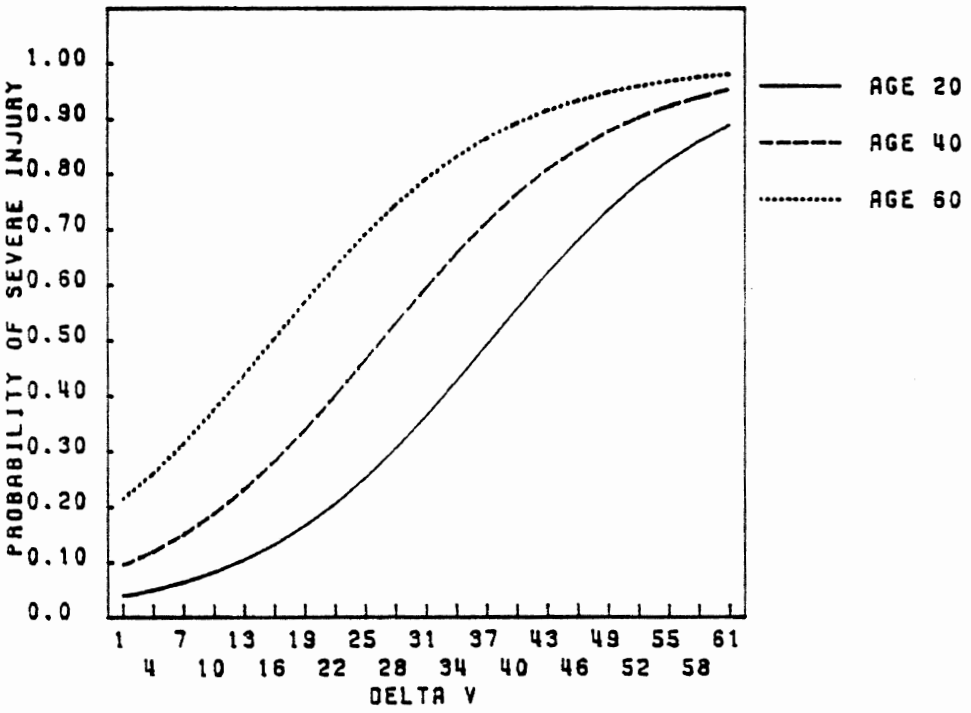

FIGURE 3.106 The Age Effect of Two-Variable Model (Delta V, Age) For OID-1VEH Phase 2 Data - Front Impacts

the Phase 1 four-variable models (Delta V, Age, No Intrusion/ Intrusion, Rural/Urban) to the Phase 2 data.

2. The Phase 2 data two-variable models represented by Equations 3-86 to 3-91 were predicting the severe injuries, better than when applying the Phase 1 two-variable models (Delta V, Age) to the Phase 2 data. The prediction of the overall injuries of these Phase 1 and Phase 2 models were not appreciably different.

The foregoing analyses indicated that, in general, the Phase 1 models (either the two-variable models or the four-variable models) predicted injury severity in the Phase 2 data nearly as well for both non-severe and severe injuries. The estimated two-variable models of both phases produced the goodness of fit results quite similarly. As with the side impact models, they showed slightly more variability in predicting the severe injuries than in predicting the non-severe injuries. But this could be caused by the fact that the sample sizes of

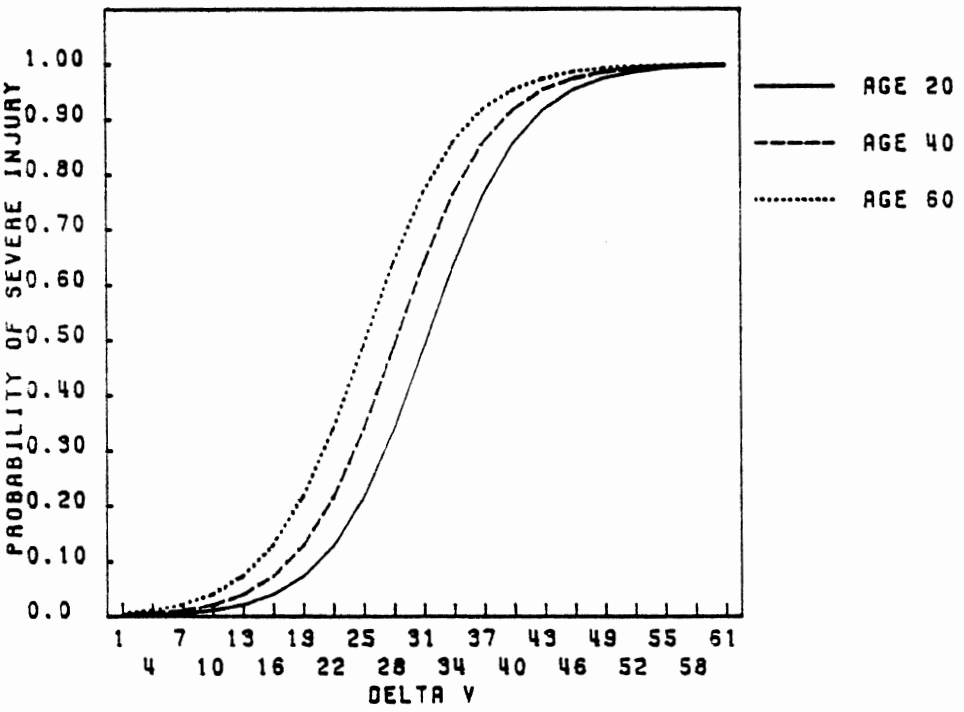

FIGURE 3.107 The Age Effect of Two-Variable Model (Delta V, Age) For OID-2VEH Phase 2 Data - Front Impacts

former were relatively smaller than the latter and that prediction of severe injuries had been, to date, somewhat tenuous. The modelling results of the Phase 1 and the Phase 2 data strongly implied that both sets of data could be combined.

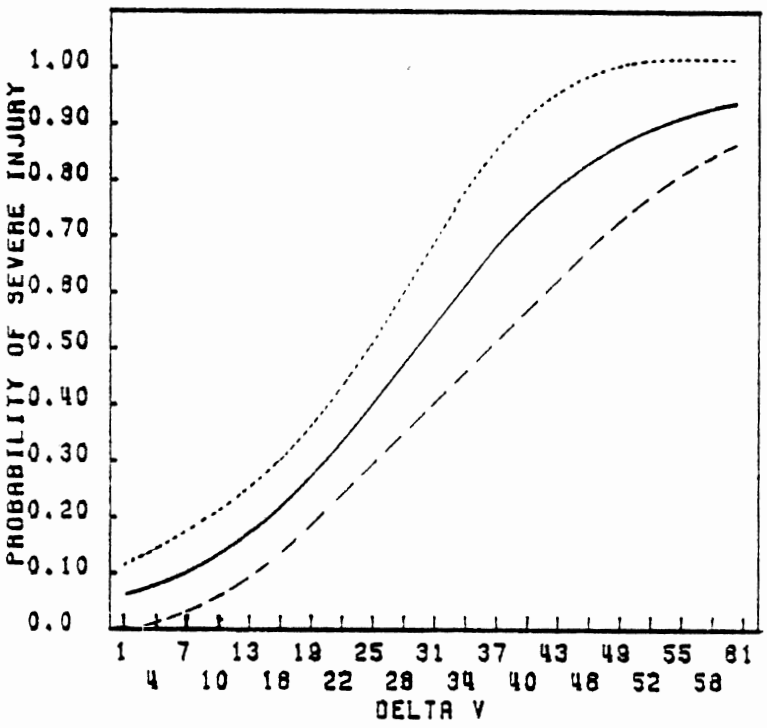

FIGURE 3.108 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) At Age 30 For CIA-1VEH Phase 2 Data - Front Impacts

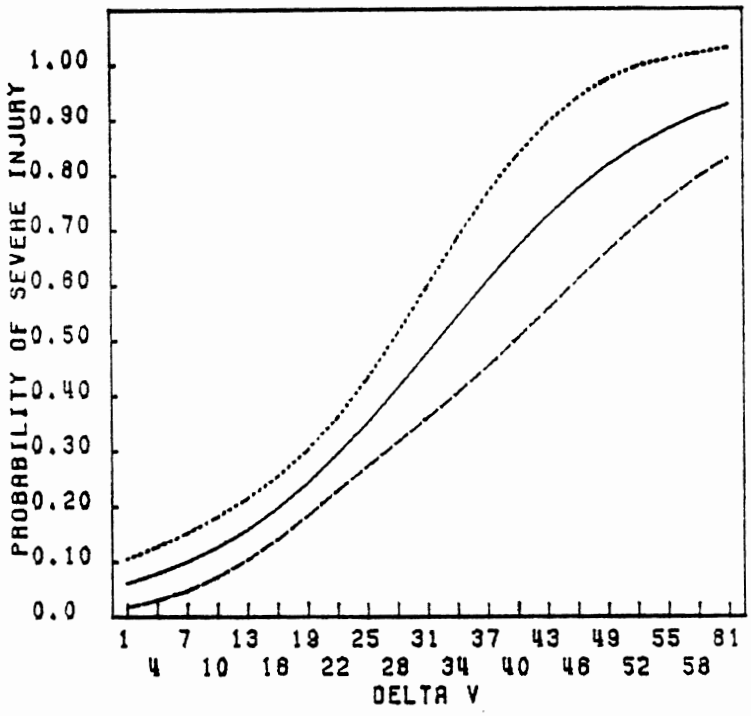

FIGURE 3.109 Confidence Interval of $\tilde{p}_i$ of Two-Variable Model (Delta V, Age) At Age 30 For OID-1VEH Phase 2 Data - Front Impacts

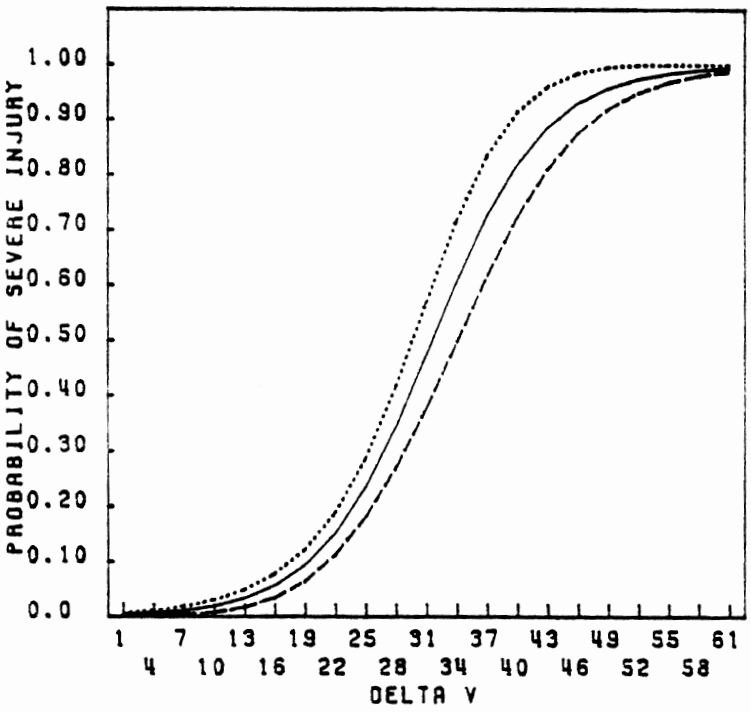

FIGURE 3.110 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) At Age 30 For CIA-2VEH Phase 2 Data - Front Impacts

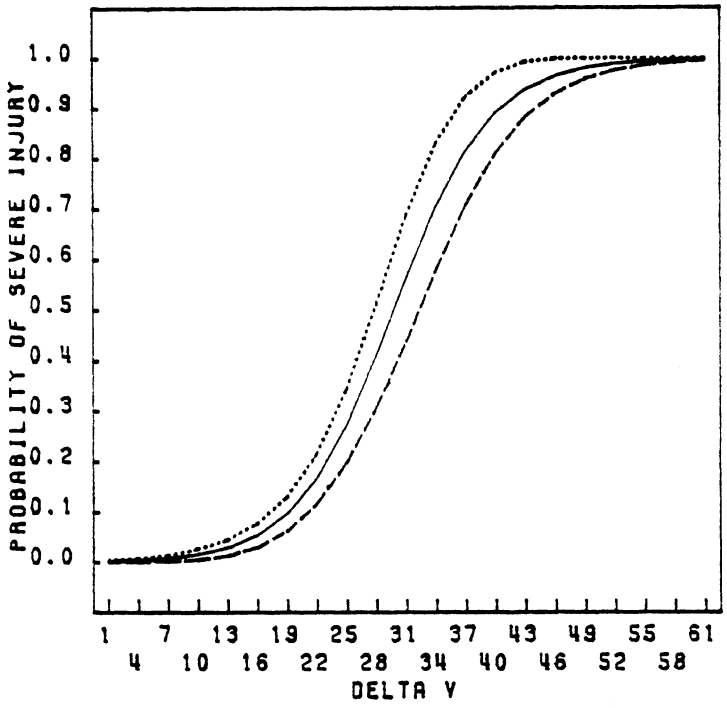

FIGURE 3.111 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) At Age 30 For OID-2VEH Phase 2 Data - Front Impacts

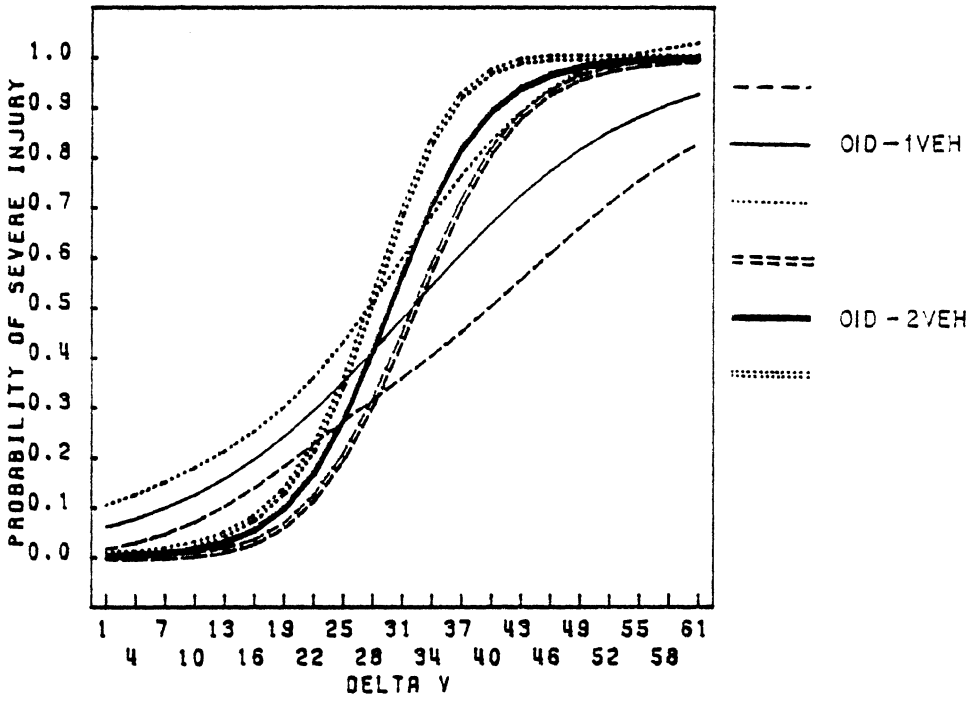

FIGURE 3.112 Confidence Intervals of $\tilde{p}_i$ of Two-Variable Model (Delta V, Age) At Age 30 For OID-1VEH and OID-2VEH Phase 2 Data - Front Impacts

TABLE 3.47

Goodness of Fit

Injury Severity = F(Delta V, Age)

Phase 2 Data - Front Impacts

Subset	Sample Size		Percentage Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
CIA-1VEH	82	41	75.6	90.2	46.3
OID-1VEH	188	66	78.3	93.6	34.8
OIP-1VEH*	98	24	81.1	96.9	16.7
CIA-2VEH	517	103	88.1	96.9	43.7
OID-2VEH	498	65	92.4	98.4	46.2
OIP-2VEH	194	35	86.0	97.4	22.9

*Age is not significant

3.5.3 Combining Phase 1 and Phase 2 Data. Further statistical investigation was carried out in order to test the plausibility of combining the Phase 1 and Phase 2 data. The statistical tests and results based on the Delta and Age models are shown in Table 3.48. The null hypothesis, H_0, is that one model will adequately describe both the Phase 1 and Phase 2 data. The alternative hypothesis, H_1, is that two independent models are required to describe the different phases. The statistical test used is the Likelihood Ratio Statistic which is discussed in more detail in Section 3.1.2. The results in Table 3.48 indicate that the Phase 1 and Phase 2 data can be combined for all frontal subsets.

TABLE 3.48

Statistical Results
Combining Phase 1 and Phase 2 Data

Subset	$-2\log L_0$[a]	$-2\log L_1$[b]	LRS[c]	df
CIA-1VEH	377.73	376.71	1.02	3
OID-1VEH	641.85	637.77	4.08	3
OIP-1VEH	241.60	236.08	5.52	3
CIA-2VEH	843.39	840.21	3.18	3
OID-2VEH	655.67	655.21	0.46	3
OIP-2VEH	388.90	384.97	3.93	3

[a] L_0 is the likelihood of the data under the null hypothesis.

[b] L_1 is the likelihood of the data under the alternative hypothesis.

[c] LRS is asymptotically chi-square with df specified.

Table 3.49 shows the proportion of severe injuries to total injuries, valid Delta V and valid Age of the six subsets for the combined data. It suggests that single-vehicle accidents are likely to result in higher proportions of severe injuries than two-vehicle accidents. Center impacts, in general, yield higher severe injury proportions than off-center impacts.

TABLE 3.49

Descriptive Statistics for Key Variables

Phases 1 and 2 - Front Impacts

Subset	Percent Severe Injuries	Delta V			Age*		
		Range	Mean	S.D.	Range	Mean	S.D.
CIA-1VEH	27.2	3-97	19.2	12.5	0-80	28.1	14.9
OID-1VEH	21.7	1-75	17.4	10.3	13-88	29.3	14.4
OIP-1VEH	19.1	2-68	16.8	9.8	0-87	23.1	14.5
CIA-2VEH	17.0	2-93	18.2	10.3	0-86	31.0	17.0
OID-2VEH	11.3	2-65	14.3	9.0	1-90	33.5	16.1
OIP-2VEH	12.9	2-57	14.9	9.2	0-90	28.0	18.9

*A zero code represents an occupant less than one year old.

Figure 3.113 shows the cumulative distribution of Delta V for the six subsets. This figure is quite similar to the cumulative distribution of Delta V for the Phase 1 data (Figure 3.67). The cumulative curves for the six subsets are quite similar.

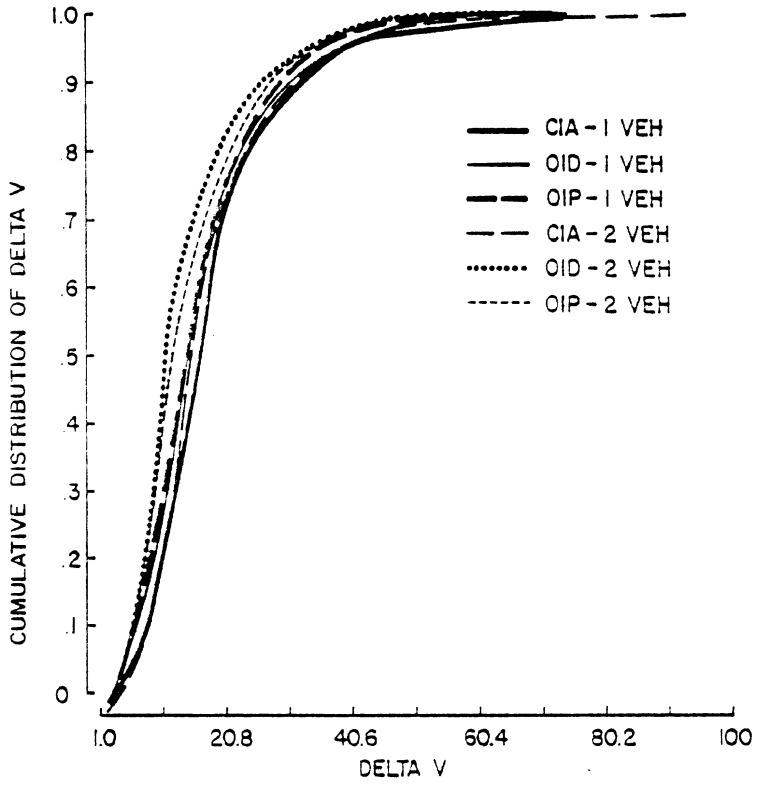

FIGURE 3.113 Cumulative Distributions of Delta V For Front-Impact Subsets Phases 1 and 2 - Front Impacts

3.5.4 Model Estimation - Phase 1 and Phase 2 Combined. The results of the modelling estimation for the combined Phase 1 and Phase 2 with Delta V and Age as the independent variables are shown in Equation 3-92 to 3-97.

Estimated Models with Delta V and Age: Phases 1 and 2

CIA-1VEH (N=366, LRS=76.92, DF=2)

(3-92) $\hat{p}_i = F(1.8051 - 0.0511X_1 - 0.0111X_2)$

OID-1VEH (N=701, LRS=125.98, DF=2)

(3-93) $\hat{p}_i = F(2.2274 - 0.0531X_1 - 0.0179X_2)$

OIP-1VEH (N=313, LRS=44.42, DF=2)

(3-94) $\hat{p}_i = F(2.1913 - 0.0483X_1 - 0.0158X_2)$

CIA-2VEH (N=1494, LRS=469.68, DF=2)

(3-95) $\hat{p}_i = F(3.5174 - 0.0912X_1 - 0.0177X_2)$

OID-2VEH (N=1588, LRS=450.64, DF=2)

(3-96) $\hat{p}_i = F(3.8541 - 0.1090X_1 - 0.0186X_2)$

OIP-2VEH (N=656, LRS=138.91, DF=2)

(3-97) $\hat{p}_i = F(2.9543 - 0.0725X_1 - 0.0197X_2)$

where

$\hat{p}_i$ is the estimated probability of a non-severe injury,
F is the logistic distribution,
X_1 is Delta V,
X_2 is Age, and
LRS is the Likelihood Ratio Statistic.

The logistic curves estimated by these equations are shown in Figure 3.114 for the six subsets. The curves show how the estimated probability of a severe injury ($1-\hat{p}_i$) varies with Delta V values holding Age fixed at 30. The three curves pertaining to the three single-vehicle subsets have similar curves which are different from the other three curves of the two-vehicle subsets. The results are similar to the Phase 1 and the Phase 2 results presented earlier. The effect of Age is

shown in Figure 3.115 and 3.116 for the OID-1VEH and OID-2VEH subsets. Each of these figures consists of three curves representing Age 20, 40 and 60. Each curve shows the variation of the estimated probability of a severe injury ($1-\hat{p}_i$) with Delta V values. In both subsets, older occupants show higher probabilities of receiving severe injuries than younger occupants for a given Delta V value. In the two-vehicle subset, the age effect approaches zero as Delta V values become very small or very large. This, however, is not quite so with the single-vehicle subset. Confidence limits for the CIA-1VEH, OID-1VEH, CIA-2VEH, and OID-2VEH subsets are shown in Figures 3.117 to 3.120. The confidence intervals were plotted holding Age fixed at 30. The bands of the two-vehicle subsets are much narrower than those of the single-vehicle subsets. For the two-vehicle subsets the confidence intervals approach zero as Delta V values become very small or very large. This is not so with the single-vehicle subsets. Comparing these with the same figures for the Phase 2 data (Figures 3.108 to 3.111) illustrates the narrowing of the confidence limits with increased sample size. Finally, the OID-1VEH and OID-2VEH subsets are shown on the same graph with their confidence limits in Figure 3.121. The figure indicates that for Delta V less than 25 mph the single-vehicle subset has higher probabilities of severe injuries than the two-vehicle subset; the reverse is true, however, for Delta V greater than 35 mph.

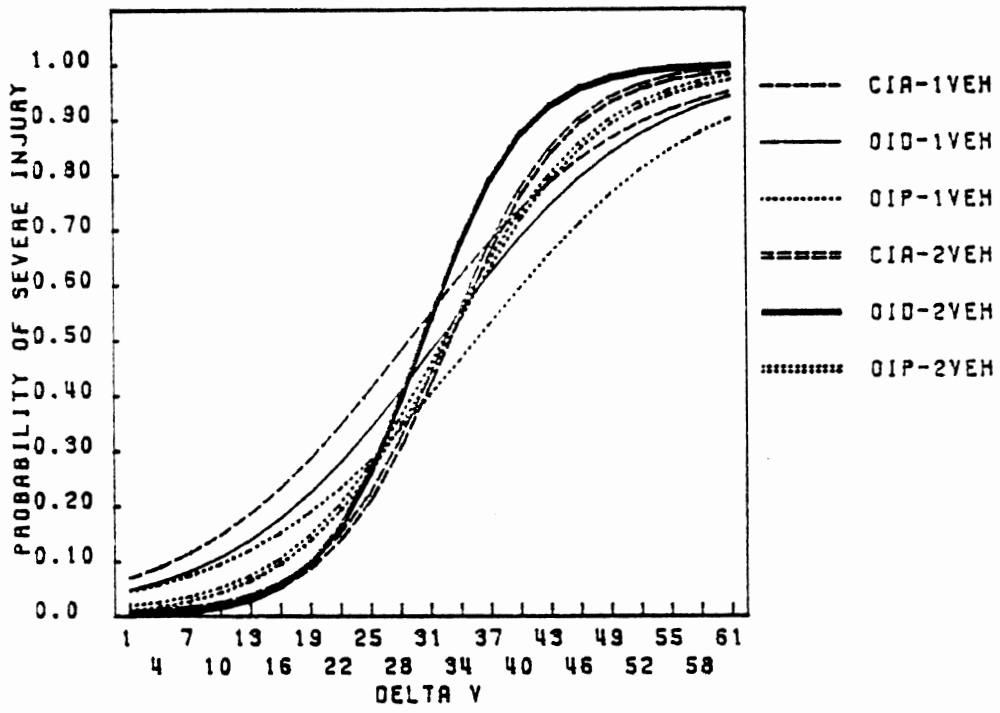

FIGURE 3.114 Logistic Curves of Two-Variable Models
(Delta V, Age) For Front-Impact Subsets
Phases 1 and 2 - Front Impacts

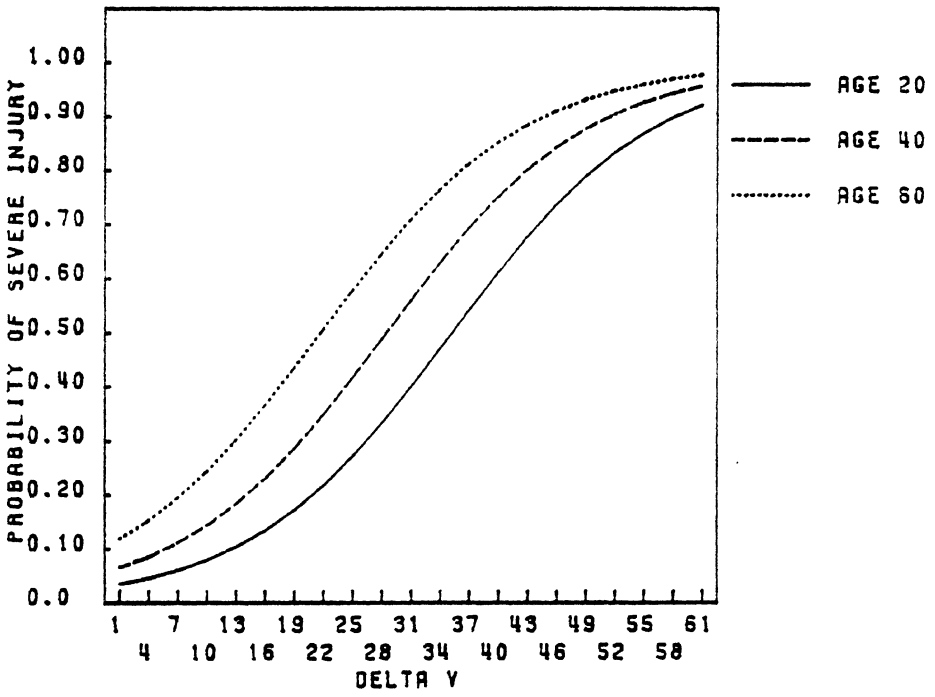

FIGURE 3.115 The Age Effect of Two-Variable Models (Delta V, Age) For OID-1VEH Phases 1 and 2 - Front Impacts

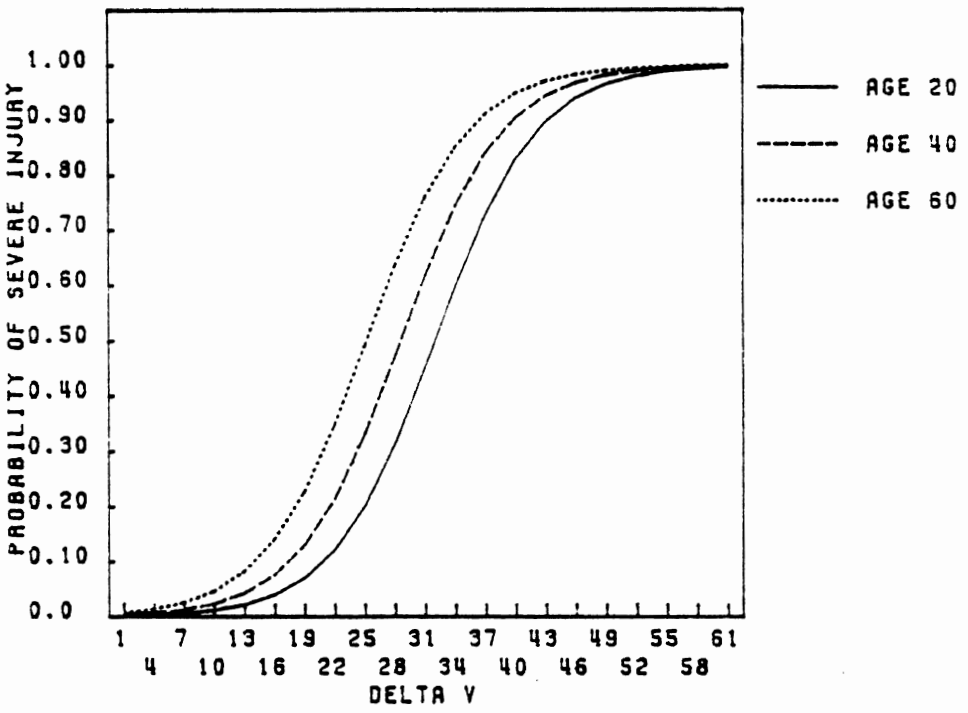

FIGURE 3.116 The Age Effect of Two-Variable Models (Delta V, Age) For OID-2VEH Phases 1 and 2 - Front Impacts

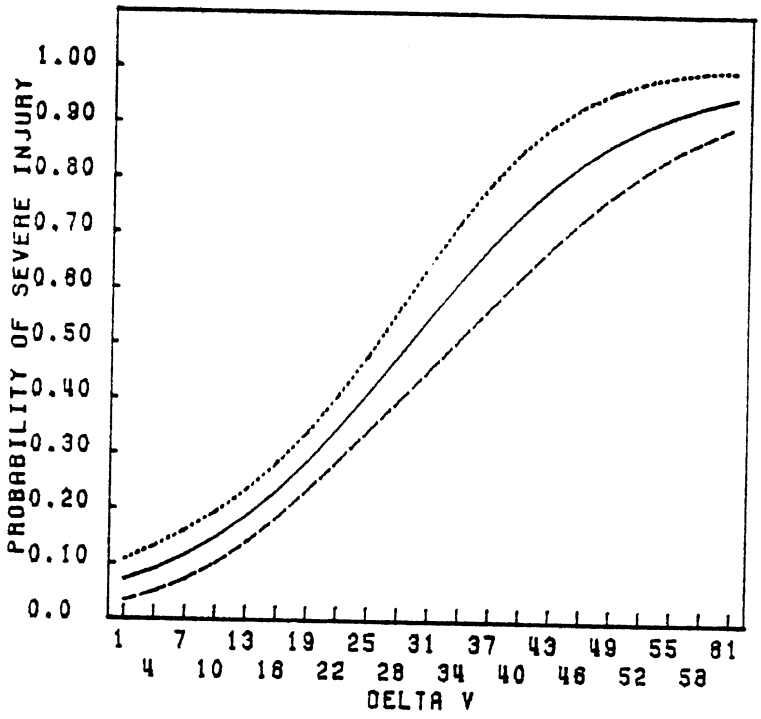

FIGURE 3.117 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For CIA-1VEH Phases 1 and 2 - Front Impacts

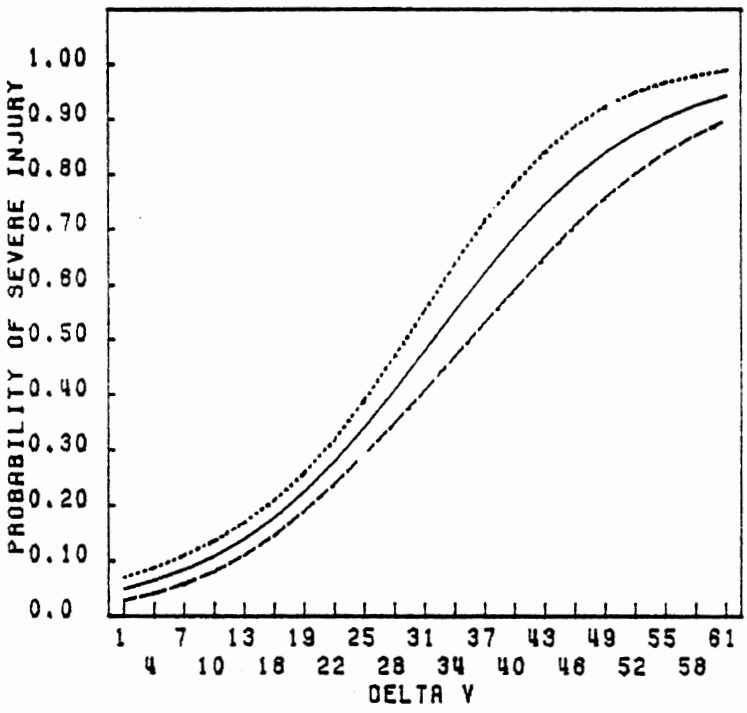

FIGURE 3.118 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For OID-1VEH Phases 1 and 2 - Front Impacts

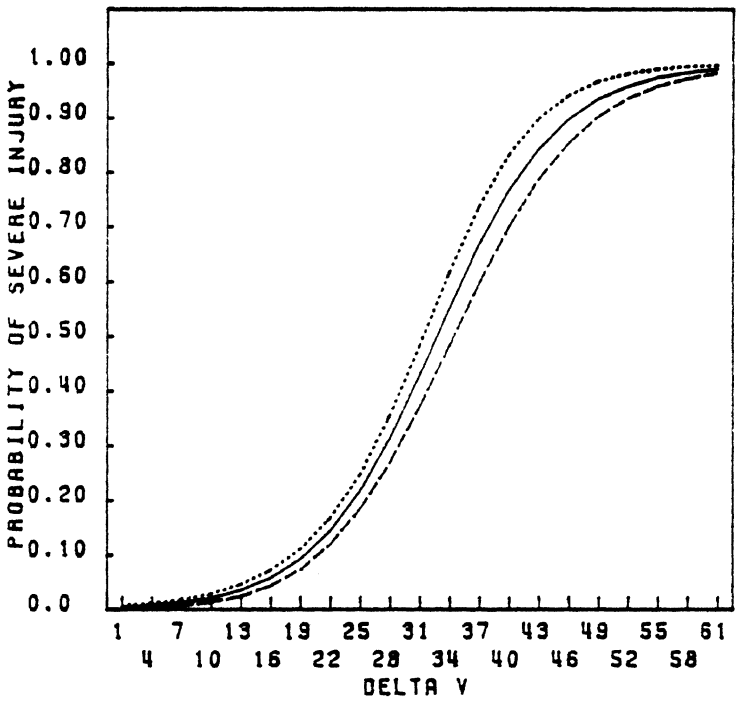

FIGURE 3.119 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For CIA-2VEH Phases 1 and 2 - Front Impacts

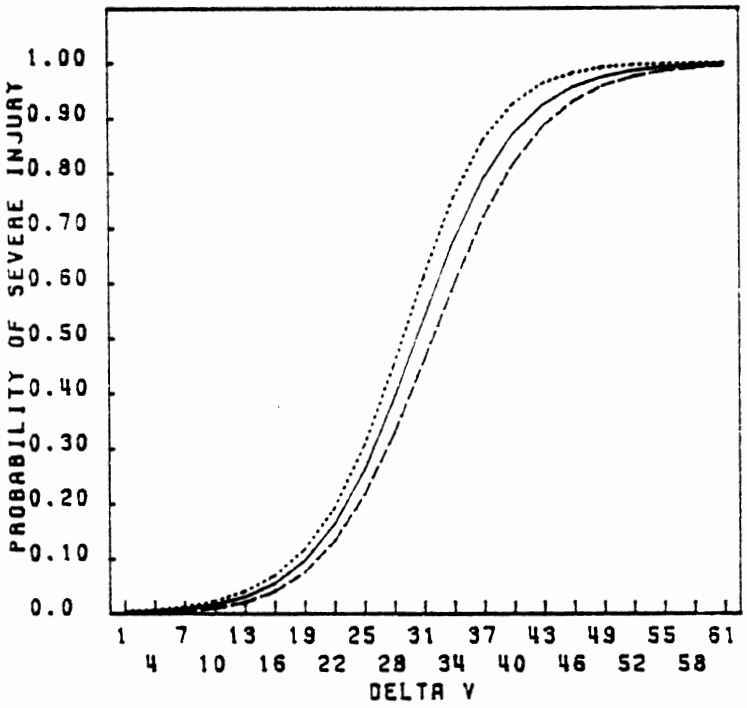

FIGURE 3.120 Confidence Interval of $\tilde{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For OID-2VEH Phases 1 and 2 - Front Impacts

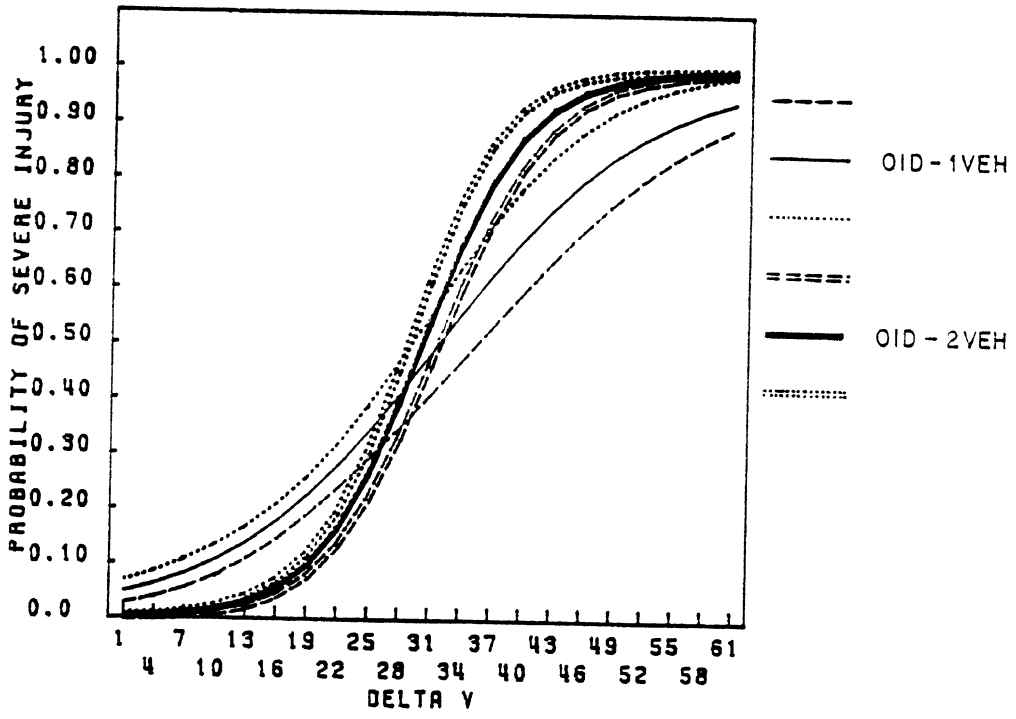

FIGURE 3.121 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For OID-1VEH and OID-2VEH Phases 1 and 2 - Front Impacts

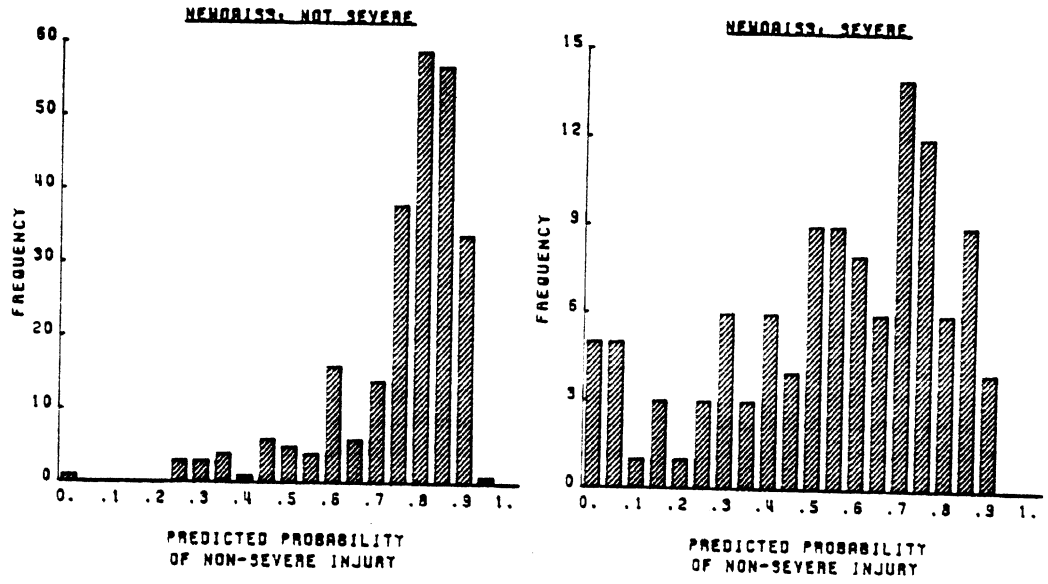

FIGURE 3.122 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For CIA-1VEH Phases 1 and 2 - Front Impacts

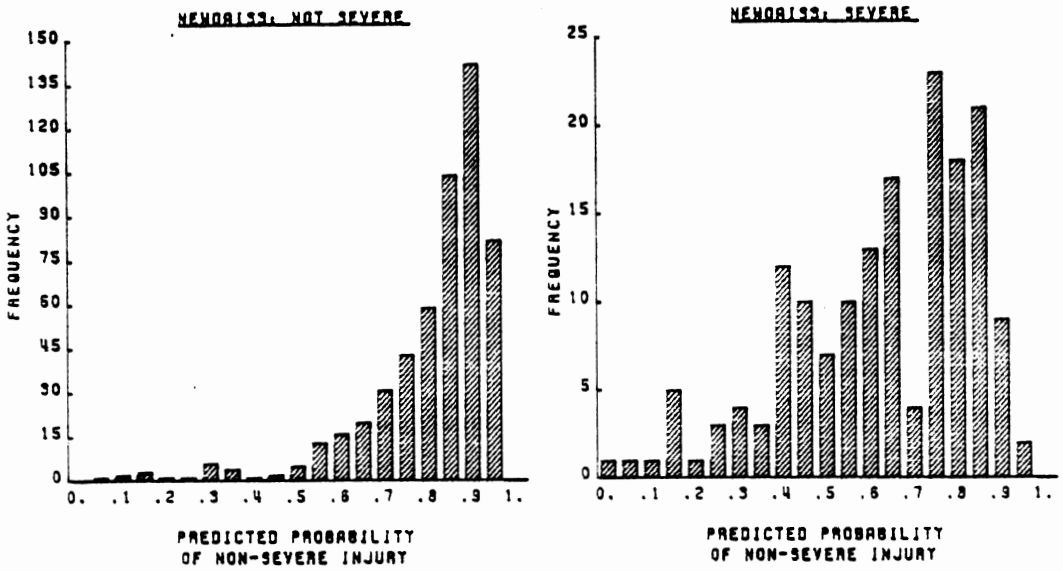

FIGURE 3.123 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OID-1VEH Phases 1 and 2 - Front Impacts

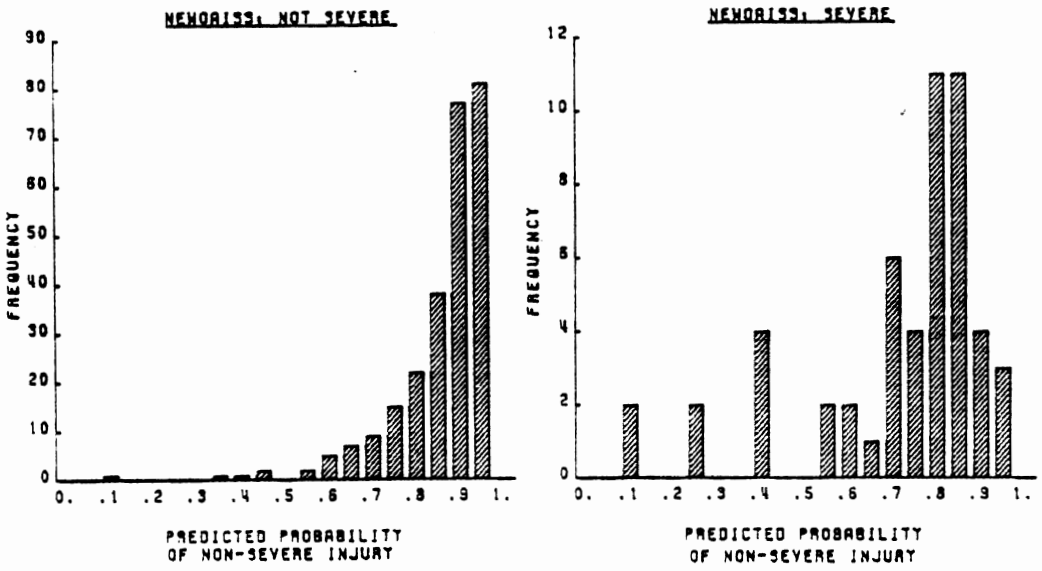

FIGURE 3.124 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OIP-1VEH Phases 1 and 2 - Front Impacts

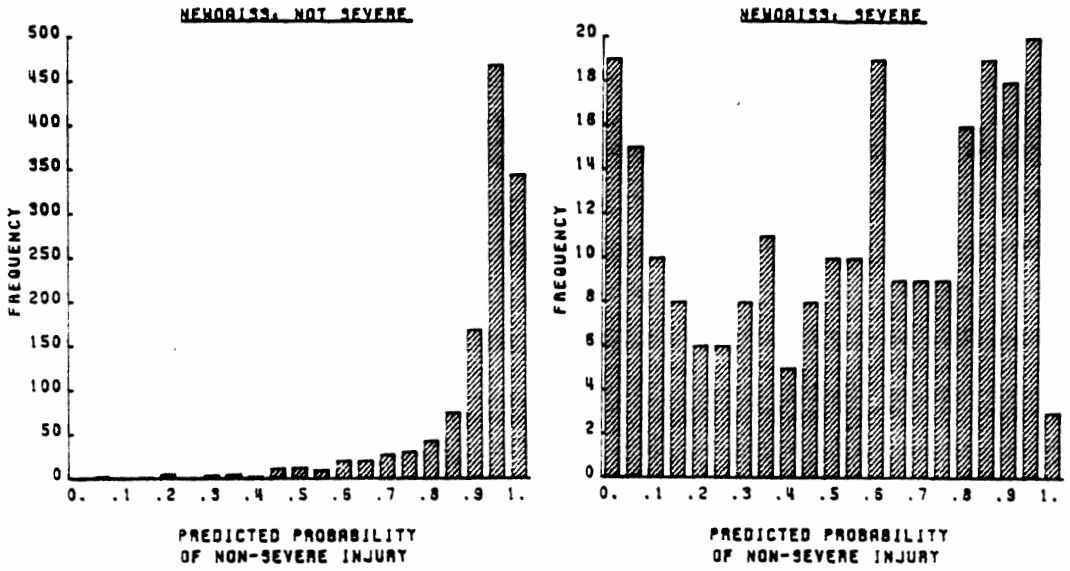

FIGURE 3.125 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For CIA-2VEH Phases 1 and 2 - Front Impacts

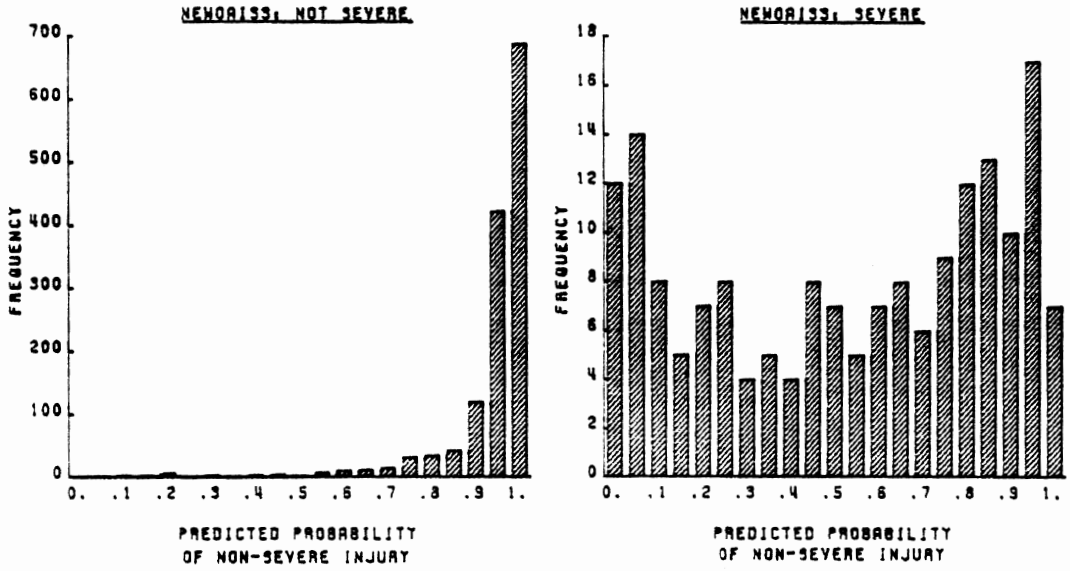

FIGURE 3.126 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OID-2VEH Phases 1 and 2 - Front Impacts

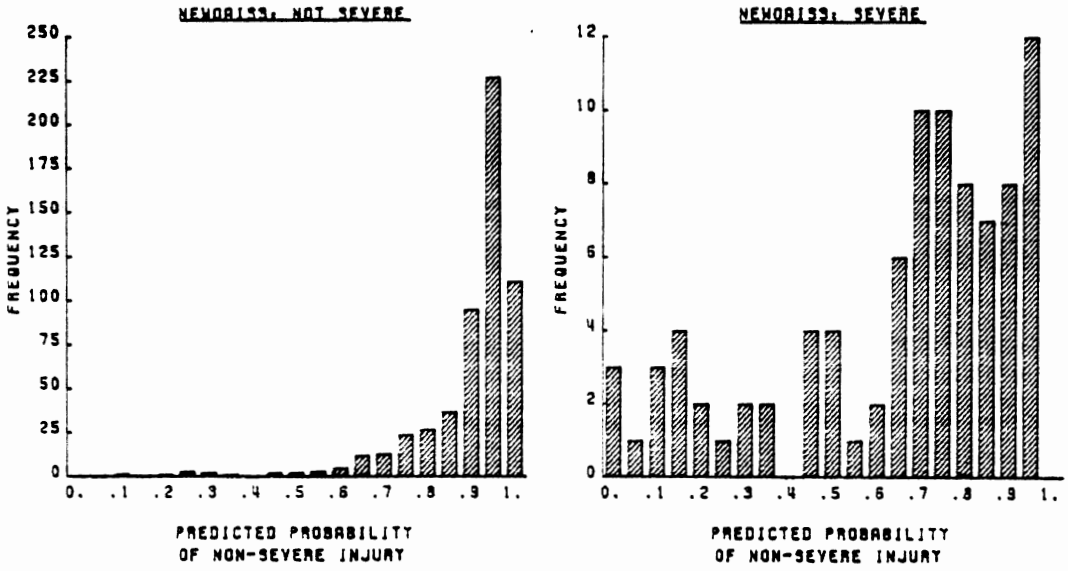

FIGURE 3.127 Histograms of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) For OIP-2VEH Phases 1 and 2 - Front Impacts

3.5.5 Model Evaluation - Phase 1 and Phase 2 Combined. The goodness of fit of the models represented by Equations 3-92 to 3-97 is shown in Table 3.50. The histograms of the $\hat{p}_i$ values for the six subsets are shown in Figures 3.122 to 3.127. The combined Phase 1 and Phase 2 two-variable models predicted the non-severe injuries most satisfactorily and consistently across all subsets. The percent correct prediction of severe injuries by these models was much lower and with more variability. The implication was that Delta V and Age alone were not quite adequate in describing a severe injury with a low or moderate value of Delta V. Other variables were needed to be investigated in order to improve the models' predictive capability, particularly those that might help describe the severe injuries. The first step was to examine variables such as Restraint Usage, Ejection, Body Region and Injury Type for the mispredicted cases.

TABLE 3.50

Goodness of Fit

Severity = F(Delta V, Age)

Phases 1 and 2 - Front Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
CIA-1VEH	252	114	74.8	92.5	36.0
OID-1VEH	536	165	79.9	95.9	27.9
OIP-1VEH	261	52	84.3	98.1	15.4
CIA-2VEH	1256	238	88.1	96.9	41.6
OID-2VEH	1412	176	92.4	98.4	44.3
OIP-2VEH	566	90	88.3	98.2	25.6

Restraint Usage. The majority of the occupants were reported not using any forms of occupant restraints. A small percentage of occupants did not have restraints while another handful used either "lap and torso" or "lap only" restraint. For the six subsets, the number of occupants using/not using restraints for cases with valid Delta V and Age is shown in Table 3.51. From Table 3.51 it could be

determined that the proportions of severe injuries to total injuries for the above four classes of restraint usage differed only slightly from one another (17% for not-used, 14% for no-restraint and 12% for lap/torso and lap only).

Table 3.52 shows the misprediction rates of severe and non-severe injuries by Restraint Usage for cases with valid Delta V and Age. Of particular interest would have been how much the rates varied within each subset across the classes of Restraint Usage for severe injuries. Unfortunately, the very small sample size of occupants using restraints tended to reduce the merits of the results shown.

Ejection. The majority of occupants were associated with no ejection or no entrapment while another handful of occupants were found either trapped or ejected or partially ejected and trapped. These proportions of occupants are shown in Table 3.53, which also reveals that the proportions of severe injuries to total injuries were appreciably higher in the presence of ejection and/or entrapment. From Table 3.53 it could be determined that for no ejection the proportion of severe injuries to total injuries was 13% while those with ejection was 63% and those with entrapment was 82%.

Table 3.54 shows the misprediction rates of severe and non-severe injuries by ejection types. For severe injuries; the proportion of cases mispredicted when the injuries involved ejection and entrapment appeared lower than when no ejection/entrapment was involved. Again, the rather small sample size of occupants with ejection/entrapment might reduce the merits of the results shown somewhat.

Table 3.55 lists the injury types which yielded the misprediction of severe injuries most frequently as well as the frequency of such misprediction for all subsets. Rupture, dislocation, hemorrhage and fracture were consistently found to be major sources of misprediction. Table 3.56 lists the affected body regions found to be associated with most frequent injury misprediction. The more common body regions in all six subsets appeared to be abdomen, chest, pelvic/hip and lower limbs. Tables 3.58 and 3.59 cross tabulate the injury types and the affected body regions for single-vehicle and two-vehicle accidents. Also tabulated were the proportion of a severe injury, the chance of it being

TABLE 3.51

Proportion of Restraint Usage for
Severe and Non-Severe Injuries

Subset	Severe Injuries					Non-Severe Injuries				
	Sample Size	Not Used (%)	No Restraint (%)	Lap/Torso (%)	Lap (%)	Sample Size	Not Used (%)	No Restraint (%)	Lap/Torso (%)	Lap (%)
CIA-1VEH	110	87	9	2	2	248	92	3	3	2
OID-1VEH	159	94	3	1	1	518	89	5	3	3
OIP-1VEH	51	90	8	0	2	251	88	9	2	1
CIA-2VEH	227	92	3	3	2	1212	89	4	4	3
OID-2VEH	171	93	1	2	3	1327	88	3	5	5
OIP-2VEH	88	92	6	1	1	549	88	9	1	2
TOTAL	806	92	4	2	2	4105	88	5	3	3

TABLE 3.52

Misprediction Rate By Restraint Usage for
Severe and Non-Severe Injuries

Phases 1 and 2 - Front Impacts

	Severe Injuries					Non-Severe Injuries				
Subset	Sample Size	Not Used (%)	No Restraint (%)	Lap/Torso (%)	Lap (%)	Sample Size	Not Used (%)	No Restraint (%)	Lap/Torso (%)	Lap (%)
CIA-1VEH	110	65	60	100	100	248	7	0	20	0
OID-1VEH	159	73	80	100	100	518	4	0	0	6
OIP-1VEH	51	83	100	NA	100	251	2	0	0	0
CIA-2VEH	227	59	63	50	75	1212	3	2	2	0
OID-2VEH	171	57	50	33	60	1327	2	2	2	2
OIP-2VEH	88	75	80	100	100	549	2	2	0	0
TOTAL	806	65	71	64	80	4105	3	1	3	2

TABLE 3.53

Proportion of Ejection/Entrapment for
Severe and Non-Severe Injuries

Phases 1 and 2 - Front Impacts

Subset	Severe Injuries					Non-Severe Injuries				
	Sample Size	No Ejection (%)	Ejection (%)	Ejection/ Trapped (%)	Trapped (%)	Sample Size	No Ejection (%)	Ejection (%)	Ejection/ Trapped (%)	Trapped (%)
CIA-1VEH	107	78	5	NA[a]	17	251	97	2	NA	1
OID-1VEH	154	77	8	1	14	527	98	1	NA	1
OIP-1VEH	49	72	14	NA	14	257	98	1	NA	1
CIA-2VEH	215	83	3	NA	14	1251	99	L.T.1[b]	NA	L.T.1
OID-2VEH	159	78	3	1	18	1398	99	L.T.1	NA	L.T.1
OIP-2VEH	86	94	2	NA	4	560	99	L.T.1	NA	L.T.1
TOTAL	770	81	5	0	14	4244	99	L.T.1	NA	L.T.1

[a]NA means no cases were recorded.

[b]L.T.1 means less than 1.

TABLE 3.54

Misprediction Rate By Ejection/Entrapment for
Severe and Non-Severe Injuries

Phases 1 and 2 - Front Impacts

Subset	Severe Injuries					Non-Severe Injuries				
	Sample Size	No Ejection (%)	Ejection (%)	Ejection/ Trapped (%)	Trapped (%)	Sample Size	No Ejection (%)	Ejection (%)	Ejection/ Trapped (%)	Trapped (%)
CIA-1VEH	107	74	25	NA*	50	251	7	25	NA	0
OID-1VEH	154	82	69	100	19	527	3	25	NA	33
OIP-1VEH	49	94	43	NA	86	257	2	0	NA	0
CIA-2VEH	215	71	0	NA	23	1251	3	75	NA	20
OID-2VEH	159	66	40	0	29	1398	2	0	NA	0
OIP-2VEH	86	78	50	NA	33	560	2	0	NA	0
TOTAL	770	75	42	50	33	4244	2	22	NA	13

*NA means no cases were recorded.

mispredicted, the associated Delta V values, the occupants' age and the corresponding contact points. The combinations of body regions and injury types in Tables 3.58 and 3.59 generally showed low to medium Delta V values. The injuries listed were usually those in which the occupants tended to come into contact with front-panels or steering assemblies of the vehicles. Other contact points were much less common although by no means insignificant. For example, injuries to Ankle/Foot were often caused by Floor/Floor Controls, and injuries to Neck were often caused by No Contact.

Further investigation on the modelling results and the outliers of the combined Phase 1 and Phase 2 data indicated that the six subsets could be combined. Based on the models with Delta V and Age as the independent variables, it was found that the three subsets of the single-vehicle accidents could be combined to form one subset and that the three subsets of the two-vehicle accidents did not appear combinable. The statistical tests for collapsing the subsets are shown in Table 3.60 and Table 3.61.

The models for these four newly defined subsets which had Delta V (X_1) and Age (X_2) as the independent variables are as follows:

Single-Vehicle Accidents (N=1380, LRS=250.29, DF=2)

(3-98) $\hat{p}_i = F(2.1130 - 0.0518X_1 - 0.0160X_2)$

Two-Vehicle Accidents (See Equations 3-95 to 3-97)

The goodness of fit results of these four models are shown in Table 3.62.

The estimated logistic curve given by Equation 3-98 for the combined single-vehicle subset is shown in Figure 3.128. The figure consists of three logistic curves, each is plotted with Age being held fixed at 30. The three curves represent the upper bound, the lower bound and the estimated probabilities of severe injuries ($1-\hat{p}_i$) of the single-vehicle subset. The confidence intervals do not approach zero when Delta V values become very small or very large although they all are somewhat reduced due to the increased sample size from the combination of three subsets. The effect of Age in this new subset is illustrated in Figure 3.129. The figure indicates that older occupants

TABLE 3.55

List of Injury Types Which Yielded A
Large Proportion of Misprediction

Phases 1 and 2 - Front Impacts

Subset	Injury Type*	Total Number of Cases	Number of Severe Injuries	Number of Mispredicted Severe Injuries
CIA-1VEH	Rupture	6	6	4
	Dislocation	9	8	3
	Fracture	79	39	34
OID-1VEH	Hemorrhage	2	2	2
	Dislocation	16	14	12
	Rupture	7	7	5
	Fracture	131	65	50
OIP-1VEH	Rupture	2	2	2
	Dislocation	6	4	4
	Fracture	74	33	28
CIA-2VEH	Hemorrhage	2	2	2
	Rupture	5	5	3
	Dislocation	15	13	10
	Fracture	202	102	77
OID-2VEH	Rupture	6	6	5
	Hemorrhage	1	1	1
	Dislocation	21	19	13
	Fracture	132	72	48
OIP-2VEH	Hemorrhage	2	2	2
	Rupture	3	3	2
	Dislocation	8	8	7
	Fracture	102	50	40

*Subject was ranked within each subset by the larger proportion of severe injuries.

have higher probabilities of receiving severe injuries than younger occupants.

Based on the new four subsets defined above, further modelling was carried out by adding the following independent variables to the models with Delta V and Age:

TABLE 3.56

List of Body Regions Which Yielded
a Large Proportion of Misprediction

Phases 1 and 2 - Front Impacts
Single-Vehicle Subsets

Subset	Body Region*	Total Number of Cases	Number of Severe Injuries	Number of Mispredicted Severe Injuries
CIA-1VEH	Pelvic/Hip	6	6	3
	Neck	5	5	4
	Lower Leg	4	4	3
	Lower Extremities	2	2	2
	Abdomen	15	14	9
	Chest	32	25	19
	Thigh	3	2	1
	Forearm	5	3	3
	Ankle/Foot	11	4	4
OID-1VEH	Abdomen	20	15	13
	Pelvic/Hip	14	10	8
	Chest	52	37	29
	Upper Arm/Forearm	20	10	8
	Ankle/Foot	22	11	9
	Lower Leg	9	4	3
	Thigh	18	9	5
OIP-1VEH	Abdomen	6	6	6
	Lower Extremities	2	2	2
	Upper Extremities	2	2	2
	Pelvic/Hip	6	4	4
	Thigh	11	5	5
	Neck	8	4	3
	Forearm	6	3	2

*Body Regions were ranked within each subset by the larger proportion of severe injuries.

Direction of Principal Force
Vehicle Weight
Vehicle Model Year
Object Contacted
Rural/Urban
Ejection
Restraint Usage
Contact Point, and

TABLE 3.57
List of Body Regions Which Yielded
a Larger Proportion of Misprediction
Two-Vehicle Subsets
Phases 1 and 2 - Front Impacts

Subset	Body Region*	Total Number of Cases	Number of Severe Injuries	Number of Mispredicted Severe Injuries
CIA-2VEH	Abdomen	24	22	13
	Pelvic/Hip	15	10	7
	Chest	101	64	40
	Thigh	19	10	7
	Forearm	18	7	5
OID-2VEH	Abdomen	27	19	14
	Pelvic/Hip	16	12	8
	Ankle/Foot	18	8	6
	Chest	75	39	24
	Forearm	19	7	4
OIP-2VEH	Abdomen	6	5	3
	Pelvic/Hip	12	7	6
	Chest	22	16	13
	Thigh	12	7	6
	Forearm	13	8	5
	Upper Arm	9	3	3

*Body Regions were ranked by the percentage of severe injuries to total injuries: The largest percentage ranked first.

Body Region

Principal Direction of Force. Principal Direction of Force was brought into the model as a set of dummy variables in the presence of Delta V and Age. It did not appear to be statistically significant, nor did it appear to significantly improve the existing models predictive capability.

Vehicle Weight. Vehicle Weight was brought into the modelling in the presence of Delta V. It did not appear to be statistically significant.

TABLE 3.58

Combination of Injury Types and Body Regions Incurring Severe Injuries
Which Were Not Correctly Predicted By The Models
(Single-Vehicle Accidents)

Phases 1 and 2 - Front Impacts

Injury Type	Body Region	Chance of Severe Injury (%)	Chance of Severe Injury Being Mispredicted (%)	Number of Severe Injuries	Delta V of Mis-predicted Cases	Age of Mis-predicted Cases	Contact Point
Rupture	Abdomen	100	69	13	14-32	15-65	steering, front-panel*
	Chest	100	100	2	17-24	22-32	steering
Hemorrhage	Chest	100	100	2	3-28	23-38	steering
Dislocation	Pelvic/Hip	100	71	14	13-26	17-50	steering, front-panel
	Wrist	100	100	2	10-20	25-26	steering, front-panel
	Ankle/Foot	86	83	6	6-27	17-32	floor/foot controls
Fracture	Neck	100	67	6	15-20	21-62	none, steering
	U.Ext/Elbow	100	100	4	14-29	18-30	-
	Chest	75	90	30	7-26		steering, front-panel
	L.Extremities	100	100	4	23-30	1-28	-
	L.Leg	92	64	11	21-27	16-59	front-panel, floor
	Pelvic/Hip	60	83	6	13-28	15-40	steering, front-panel
	Ankle/Foot	50	92	12	9-24	13-57	floor/foot controls
	Knee	70	86	7	8-28	16-37	front-panel
	Thigh	53	69	16	8-33	15-43	front-panel, steering
	Forearm	71	75	12	10-25	14-63	instrument panel, steering
Laceration	Abdomen	100	87	15	4-29	18-57	instrument panel steering
Contusion	Abdomen	60	83	6	11-30	16-25	steering

*Front-panel includes instrument panel and glove compartment.

TABLE 3.59

Combination of Injury Types and Body Regions Incurring Severe Injuries
Which Were Not Correctly Predicted By The Models
(Two-Vehicle Accidents)

Phases 1 and 2 - Front Impacts

Injury Type	Body Region	Chance of Severe Injury (%)	Chance of Severe Injury Being Mispredicted (%)	Number of Severe Injuries	Delta V of Mis-predicted Cases	Age of Mis-predicted Cases	Contact Point
Rupture	Abdomen	100	75	12	11-28	17-67	steering, front-panel*
	Chest	100	50	21	28	46	steering, front-panel
Dislocation	Pelvic/Hip	100	71	21	17-35	18-66	steering, front-panel
	Neck	100	80	5	12-26	17-64	none
	Ankle/Foot	67	88	8	6-31	17-73	floor/foot controls
	Wrist/Hand	100	100	2	8-10	21-24	-
Hemorrhage	Chest	100	100	3	9-20	21-66	steering
Fracture	Neck	94	59	17	10-31	15-69	none, instrument-panel
	Chest	73	86	65	3-33	16-83	steering, front-panel, side
	Thigh	62	65	23	12-37	2-63	front-panel, steering
	Lower Leg	63	84	19	16-23	2-45	front-panel
	L.Extremities	100	60	5	26-27	14-51	instrument panel
	Forearm	61	64	22	15-33	16-86	front-panel, steering
	U.Extremities	100	100	2	23-31	18-27	instrument panel
	Upper Arm	75	83	6	9-25	8-69	-
Laceration	Abdomen	100	56	25	15-32	17-60	steering, front-panel
Contusion	Abdomen	54	86	7	14-32	17-69	steering

*Front-panel includes instrument panel and glove compartment.

TABLE 3.60

Statistical Results
Combining CIA-1VEH, OID-1VEH and OIP-1VEH

Phases 1 and 2 - Front Impacts

Hypothesis	-2 Log L	df
H_0	1270.20	3
H_1	1261.28	9
	Chi-square = 8.92	6

where H_0 :the subsets have the same model

H_1 :the subsets have different models

TABLE 3.61

Statistical Results
Combining CIA-2VEH, OID-2VEH and OIP-2VEH

Phases 1 and 2 - Front Impacts

Hypothesis	-2 Log L	df
H_0	1909.90	3
H_1	1887.96	9
	Chi-square = 21.94	6

H_0 :the subsets have the same model

H_1 :the subsets have different models

Object Contacted. Object contacted was brought into the modelling in two forms: a set of dummy variables and a set of

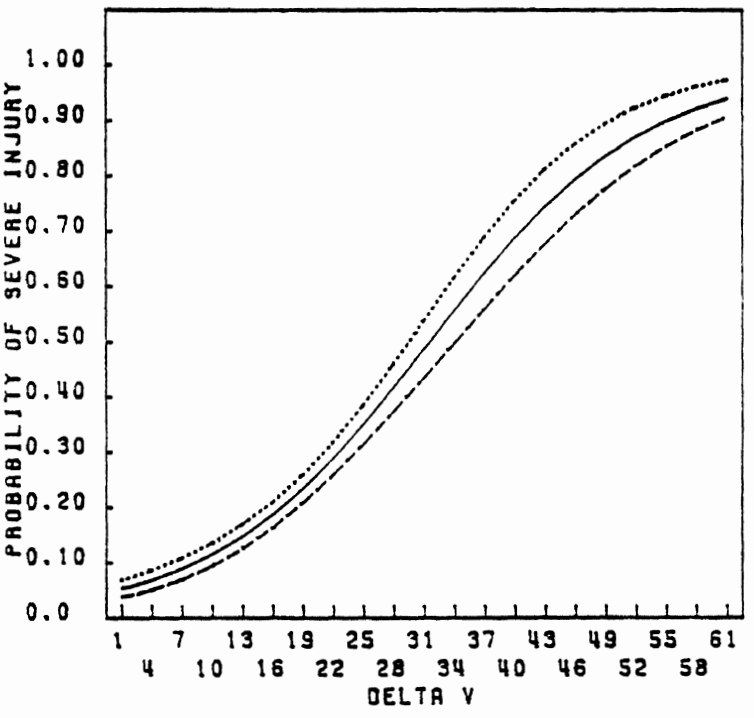

FIGURE 3.128 Confidence Interval of $\hat{p}_i$ of Two-Variable Model (Delta V, Age) at Age 30 For The Single-Vehicle Accident Subset Phases 1 and 2 - Front Impacts

interaction terms with Delta V. Neither form of Object Contacted was statistically significant in the presence of Delta V and Age. Both slightly improved the prediction of the severe injury cases .

Rural/Urban. Rural/Urban was brought into the modelling as a dummy variable. While it appeared to be a statistically significant independent variable for all the subsets, it failed to significantly improve the existing models' predictive capability.

Restraint Usage. Restraint Usage was brought into the model as a set of dummy variables; Not Used, No Restraint (older model cars), Lap/Torso and Lap Only. It did not appear to be a statistically significant independent variable.

Ejection. Ejection was brought into the models, in the presence of Delta V and Age, in the following manner:

$$\hat{p}_i = F(\text{Delta V, Age, Ejection Dummy, and Ejection-Dummy X Delta V})$$

The Ejection dummy variables were coded as follows:

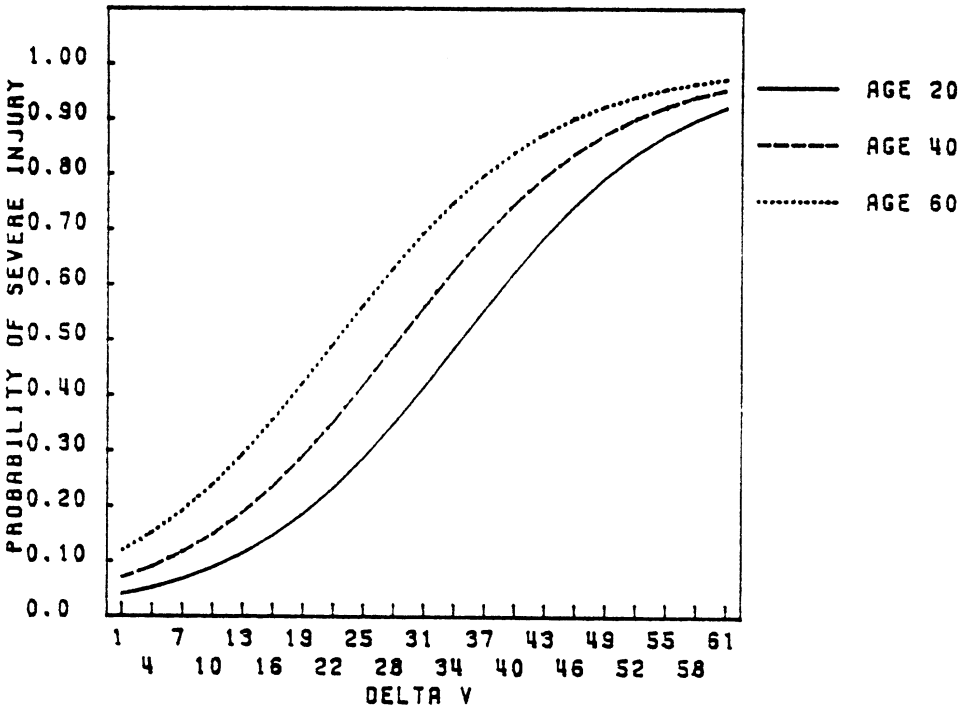

FIGURE 3.129 The Age Effect of Two-Variable Model (Delta V, Age) For The Single-Vehicle Accident Subset Phases 1 and 2 - Front Impacts

TABLE 3.62

Goodness of Fit

Severity = F(Delta V, Age)

Phases 1 and 2 - Front Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
SINGLE-VEHICLE	1049	331	79.3	95.7	27.2
CIA-2VEH	1256	238	88.1	96.9	41.6
OID-2VEH	1412	176	92.4	98.4	44.3
OIP-2VEH	566	90	88.3	98.2	25.6

e_1 = 1 if entrapped
0 otherwise

e_2 = 1 if ejected, partially ejected, ejected unknown

degree, ejected/trapped
0 otherwise

e_3 = 1 no entrapment, no ejection
0 otherwise

Ejection and the interaction of Ejection with Delta V appeared to be a significant variable in the single-vehicle subset. The model can be expressed as:

Single-Vehicle (N=1345, LRS=294.55, DF=6)

$$(3\text{-}99) \quad \hat{p}_i = F(2.2102 - 0.0482X_1 - 0.0181X_2 - 2.1774X_3 - 1.8628X_4 + 0.0333X_5 + 0.0836X_6)$$

where

X_1 is Delta V,

X_2 is Age,

X_3 is equal to 1 if there was entrapment and 0 otherwise,

X_4 is 1 if there was ejection and 0 otherwise,

X_5 is X_3 times Delta V,

X_6 is X_4 times Delta V,

$\hat{p}_i$ is the estimated probability of a non-severe injury, and

F is the logistic distribution function.

The goodness of fit of Equation 3-99 was:

Overall percent correct prediction = 81.0%
Non-Severe injury percent correct prediction = 95.1%
Severe injury percent correct prediction = 34.2%

For the two-vehicle subsets, however, ejection appeared not significant. Neither did it improve the models' predictive capability. That Ejection appeared to be a significant variable as the sample size had considerably increased reflects the underlying influence of this variable on injury prediction. In the situation where there are sufficient data, particularly on those being ejected or trapped the extent of the influence of this variable may be more meritoriously determined.

Contact Point. In frontal collisions, the contact points which occurred with any significant frequency were:

Instrument Panel
Glove Compartment

Steering Assembly
Windshield
Hardware
A-Pillar
Side Interior Surface
Floor
Foot Controls

Table 3.63 shows the number of cases with valid and missing contact point information with valid Delta V, Age and OAIS Code. If Contact Point were to be put in the modelling the number of valid cases would be reduced by more than 50%. Table 3.64 shows the proportion of severe injuries to total injuries by various contact points.

TABLE 3.63

Number of Valid and Missing Cases of Contact Point

Phases 1 and 2 - Front Impacts

Subset	Sample Size	Missing Cases
SINGLE-VEHICLE	663	714
CIA-2VEH	682	800
OID-2VEH	536	1052
OIP-2VEH	287	363

The contact point variable was brought into the modelling in two forms; a set of dummy variables and the interaction terms of contact point and Delta V. The Contact Point dummy variables were coded as follows:

d_1 = 1 if instrument panel or glove compartment
0 otherwise

d_2 = 1 if steering assembly
0 otherwise

d_3 = 1 if side interior or A-pillar
0 otherwise

d_4 = 1 if floor or floor control
0 otherwise

TABLE 3.64

Proportion of Severe Injuries by
Contact Point

Phases 1 and 2 - Front Impacts

Contact Point	Single-Vehicle		Two-Vehicle	
	Sample Size	Proportion of Severe Injuries	Sample Size	Proportion of Severe Injuries
Instrument Panel	182	30.8	393	30.3
Glove Compartment	55	43.6	108	32.4
Steering	318	39.3	520	30.2
Windshield	271	5.9	531	5.1
Hardware	12	33.3	27	22.2
A-Pillar	17	17.6	52	25.0
Side Interior Surface	13	30.8	44	36.4
Floor	34	52.9	37	21.6
Foot Controls	17	52.9	27	22.2
No Contact	43	32.6	145	14.5

d_5 = 1 if other contact point
0 otherwise

d_6 = 1 if no contact injury source
0 otherwise

Although the contact points variable (in either form) seemed to have improved the percent correct prediction of severe injuries significantly, the marked reduction in the sample size (over 50%) made the variable less desirable.

A close examination of Tables 3.58 and 3.59 revealed that for one combination of Body Region and Injury Type only one or sometimes two contact points were major causes of the injury. This seemed to suggest that Contact Point and Body Region should be good explanatory variables of injury severity in the presence of Delta V and Age. But since Contact Point was correlated with Body Region, only one of them could be included in the models. Body Region was likely to be a better variable than Contact Point because, from Tables 3.58 and 3.59, a Body Region seemed to readily suggest a certain contact point whereas a contact point could very well imply numerous different body regions.

3.5.7 Final Models. Based on the contact point distribution within each body region and the probability of the severe injury of each body region, the body region variable was structured to have the following levels:

1. Head/Skull, Neck
2. Upper Extremities, Elbow, Forearm and Wrist/Hand
3. Chest and Abdomen
4. Pelvic/Hip, Lower Extremities, Thigh and Knee
5. Ankle/Foot, Lower Leg
6. Other

A set of dummy variables was created for the above levels of Body Region and this was incorporated into the modelling in the presence of Delta V and Age. The following modelling results were obtained:

Final Estimated Models for Front Impacts: Phases 1 and 2

Single-Vehicle Accidents (N = 1380 LRS = 493.9)

(3-100) $\hat{p}_i = F(2.5013 - 0.0460X_1 - 0.0135X_2 - 0.6385X_3 - 1.1439X_4 - 1.6394X_5 - 1.4079X_6 - 1.0793X_7)$

CIA-2VEH (N = 1492 LRS = 622.1)

(3-101) $\hat{p}_i = F(3.6405 - 0.0841X_1 - 0.0092X_2 - 0.6376X_3 - 0.8305X_4 - 1.7620X_5 - 1.4753X_6 - 0.7237X_7)$

OID-2VEH (N = 1587 LRS = 565.4)

(3-102) $\hat{p}_i = F(3.9327 - 0.0998X_1 - 0.0119X_2 - 0.3985X_3 - 0.6505X_4 - 1.6476X_5 - 1.5381X_6 - 0.7334X_7)$

OIP-2VEH (N = 656 LRS = 205.3)

(3-103) $\hat{p}_i = F(3.1374 - 0.0622X_1 - 0.0184X_2 - 0.2414X_3 - 1.1911X_4 - 1.8419X_5 - 1.4230X_6 - 0.3409X_7)$

where

X_1 is Delta V,

X_2 is Age,

X_3 is 1 if a neck or head/skull injury and 0 otherwise,

X_4 is 1 if the injury is in the upper extremities, elbow, forearm, and 0 otherwise,

X_5 is 1 if injury is in the chest or abdomen and 0 otherwise,

X_6 is 1 if the injury is in the pelvic/hip area, lower extremities or thigh and 0 otherwise,

X_7 is 1 if the injury is in the ankle/foot or lower leg and 0 otherwise,

$\hat{p}_i$ is the estimated probability of a non-severe injury, and

F is the logistic distribution.

Table 3.65 shows the goodness of fit of the models represented by Equations 3-100 to 3-103. Figures 3.130 to 3.133 show the histograms of the $\hat{p}_i$ values for the four subsets. Each figure, representing a model of a particular subset, consists of two histograms, one for non-severe injuries and the other for severe injuries. By having body Region in

TABLE 3.65

Goodness of Fit

Severity = F(Delta V, Age, Body Region)

Phases 1 and 2 - Front Impacts

Subset	Sample Size		Percent Correct Prediction		
	Non-Severe	Severe	Overall	Non-Severe	Severe
SINGLE-VEHICLE	1049	331	83.4	92.4	55.0
CIA-2VEH	1254	238	90.8	96.8	59.2
OID-2VEH, Driver	1411	176	93.4	97.7	59.1
OIP-2VEH, Non-Driver	566	90	90.7	97.0	51.1

the models, the percent correct prediction of severe injuries improved most considerable.

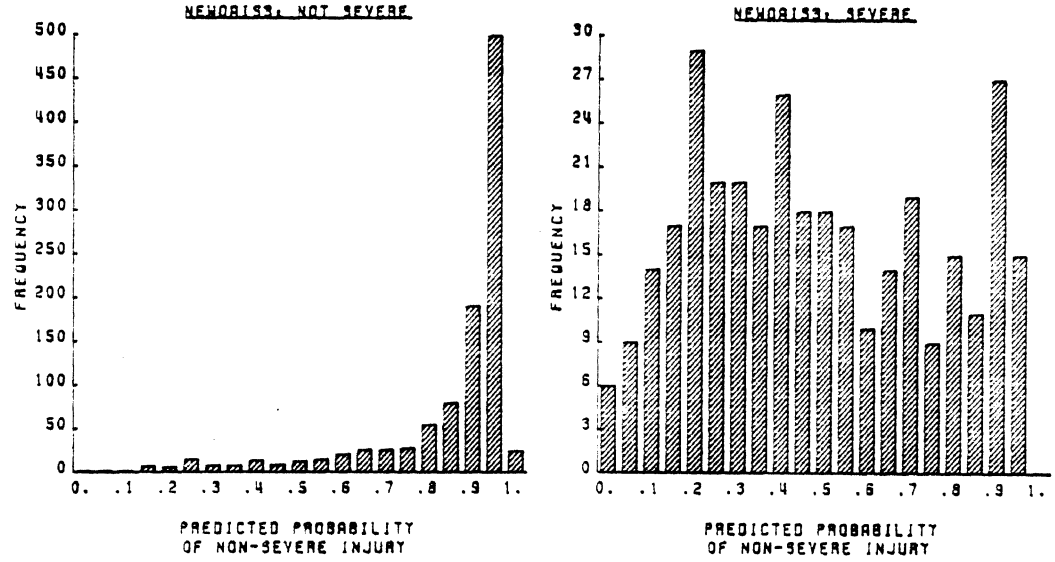

FIGURE 3.130 Histograms of $\hat{p}_i$ of Three-Variable Model (Delta V, Age, Body Region) For The Single-Vehicle Accident Subset Phases 1 and 2 - Front Impacts

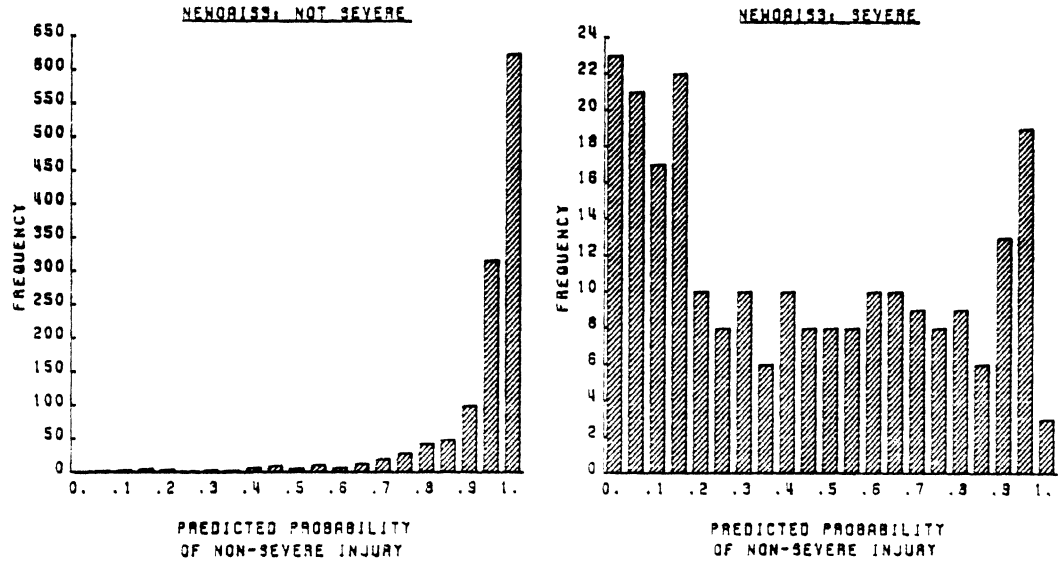

FIGURE 3.131 Histograms of $\hat{p}_i$ of Three-Variable Model (Delta V, Age, Body Region) For CIA-2VEH Phases 1 and 2 - Front Impacts

The estimated logistic curves of the final models for the single-vehicle accident subset and CIA-2VEH are shown in Figures 3.134 and 3.135. Each figure consists of three logistic curves representing the upper bound, the lower bound and the estimated probability of a severe injury ($1-\hat{p}_i$) in Upper Extremities or Elbow or Forearm as a function of Delta V with Age fixed at 30. Figure 3.135 indicates the 95% confidence

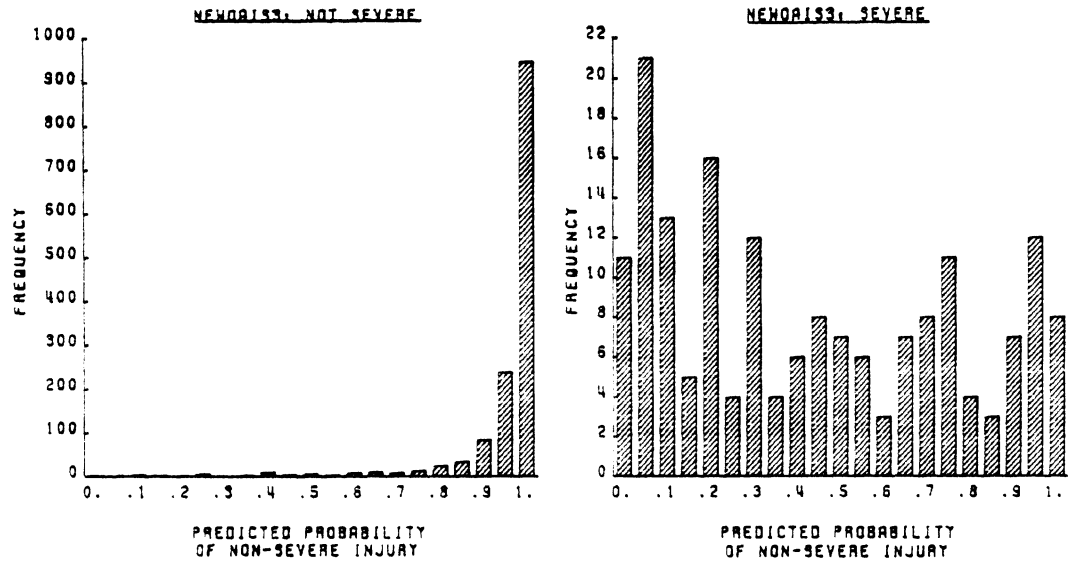

FIGURE 3.132 Histograms of $\hat{p}_i$ of Three-Variable Model (Delta V, Age, Body Region) For OID-2VEH Phases 1 and 2 - Front Impacts

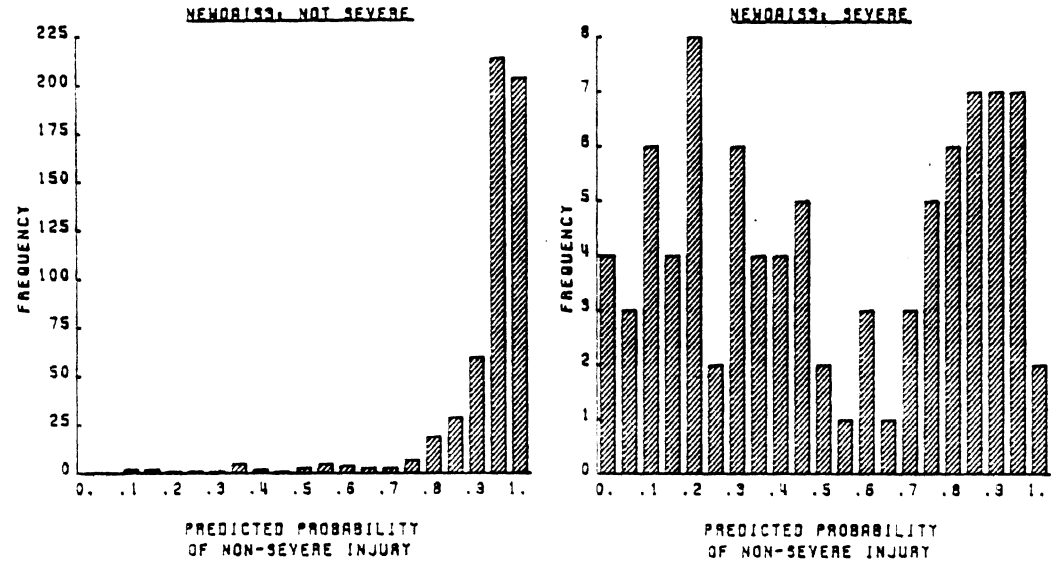

FIGURE 3.133 Histograms of $\hat{p}_i$ of Three-Variable Model (Delta V, Age, Body Region) For OIP-2VEH Phases 1 and 2 - Front Impacts

interval of the CIA-2VEH subset approaching zero as Delta V values become increasingly small or large; this is not so with the single-vehicle subset (Figure 3.134). The effects of the five levels of the Body Region dummy variable are illustrated in Figures 3.136 and 3.137 for the single-vehicle accident model and the CIA-2VEH respectively. For the single-vehicle model (Figure 3.136), the curve on the extreme right represents either the situation in which no body regions were specified as being injured or the injuries to the body regions were

other than those specified in the model; either of these tended to be associated with low proportions of severe injuries. This figure indicates that when chest or abdomen is injured the probability of a severe injury is quite high even when Delta V is low (about 10 mph). Pelvic/Hip and Lower Extremities are also likely to sustain severe injuries at the low to moderate values of Delta V (about 20 mph). Upper Extremities, Forearm, Elbow and Ankle/Foot, Lower Leg are more likely to result in severe injuries when Delta V values are greater than 25 mph. Head and Neck are prone to a severe injury when Delta V values exceed 35 mph. Injuries to other body regions or to body regions not specified are likely to be severe when Delta V values are greater than 45 mph.

For the CIA-2VEH model (Figure 3.137) the curve on the extreme right represents either the situation in which no body regions were specified as being injured or the injury to the body regions other than those specified in the model; both tended to have low proportions of severe injuries. Like the single-vehicle subset, Chest or Abdomen, Pelvic/Hip, Lower Extremities are more likely to suffer a severe injury than other body regions. The Limbs, Head and Neck show similar probabilities of receiving severe injuries. Chest and Abdomen are likely to be severely injured when Delta V exceeds 20 mph. Injuries in Pelvic/Hip, Lower Extremities and Thigh are likely to be severe for Delta V exceeding 22 mph. Upper Extremities, Elbow, Forearm, Ankle/Foot, Lower Leg, Head and Neck are likely to sustain severe injuries when Delta V exceeds 30 mph. Other body regions not mentioned above will not be likely to sustain severe injuries until Delta V is greater than 40 mph.

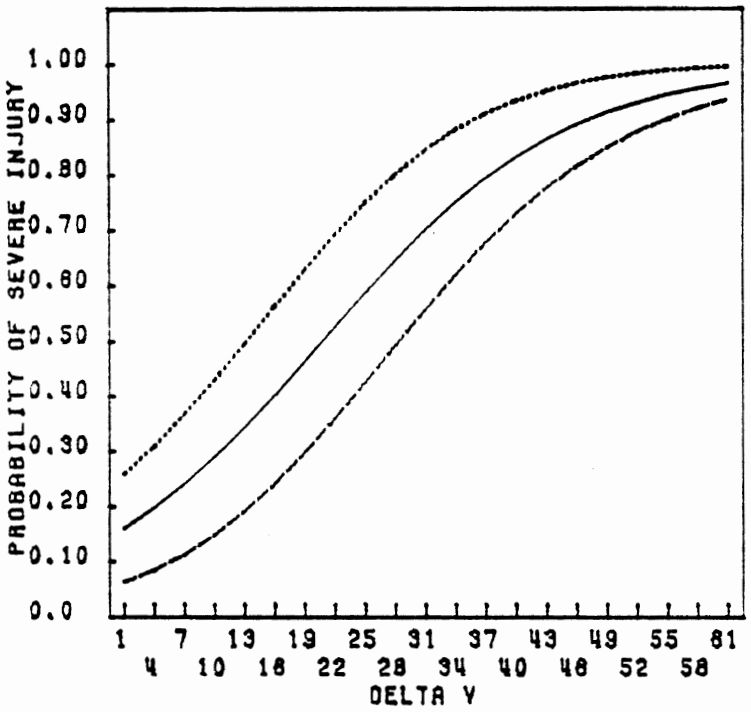

FIGURE 3.134 Confidence Interval of $\hat{p}_i$ of Three-Variable Model (Delta V, Age, Body Region) at Age 30 For The Single-Vehicle Accident Subset Phases 1 and 2 - Front Impacts

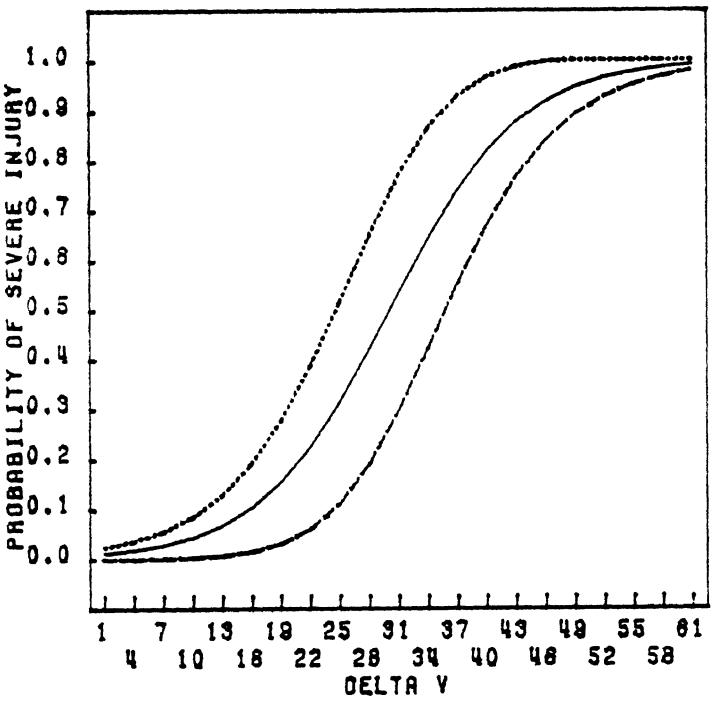

FIGURE 3.135 Confidence Interval of $\hat{p}_i$ of Three-Variable Model (Delta V, Age, Body Region) at Age 30 For CIA-2VEH Phases 1 and 2 - Front Impacts

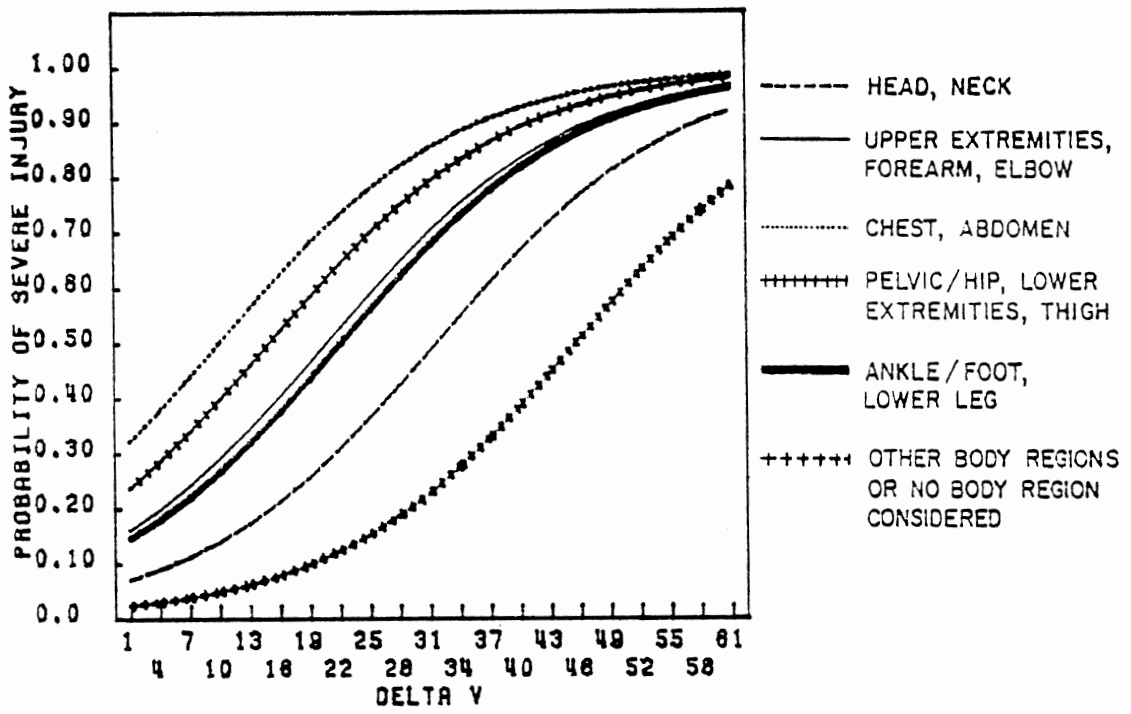

FIGURE 3.136 The Effect of Five Levels of Body Region of Three-Variable Model (Delta V, Age, Body Region) For The Single-Vehicle Accident Subset
Phases 1 and 2 - Front Impacts

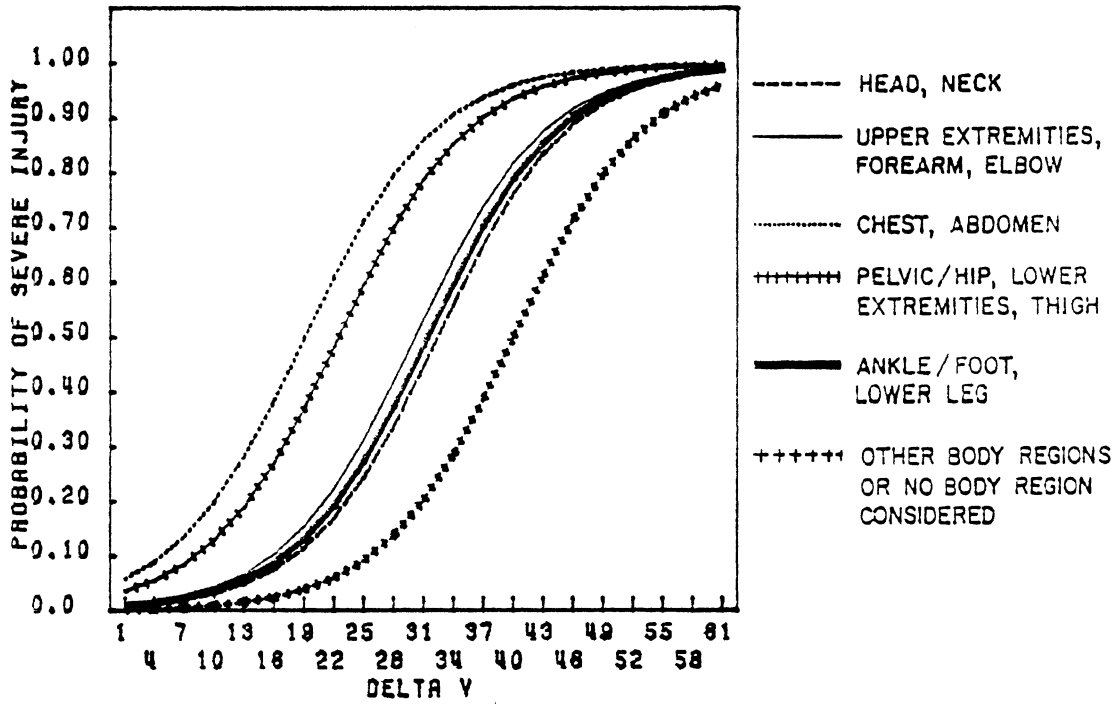

FIGURE 3.137 The Effect of Five Levels of Body Region of Three-Variable Model (Delta V, Age, Body Region) For CIA-2VEH Phases 1 and 2 - Front Impacts

3.5.8 Significant Results. The analysis of the Phase 2 data indicated that it was not different from the Phase 1 data, and that data from both phases could be combined. The Phase 1 two-variable models (Delta V and Age) fitted the Phase 2 data almost as well. Moreover, the Phase 1 two-variable models and the Phase 2 two-variable models fitted the respective data with no appreciable differences. While in the Phase 1 data the occurrence of intrusion appeared quite significant in explaining injury severity when Delta V and Age were both in the model, its influence in the Phase 2 models was almost negligible. Intrusion information in the Phase 1 and Phase 2 data was recorded quite differently. It was possible that this might have caused the discrepancy in the effect of the intrusion variable. Rural/Urban, which had shown tenuous influence in predicting injury severity in the Phase 1 data, turned out to be not significant in the Phase 2 models.

The data of both phases were combined and the six frontal impact subsets of the combined data were collapsed to form the following four independent subsets:

1. single-vehicle accidents
2. two-vehicle accidents with center impacts (CIA-2VEH)
3. two-vehicle accidents, off-center impacts, and drivers only (OID-2VEH)
4. two-vehicle accidents, off-center impacts, and right-front passengers (OIP-2VEH).

The model estimation results for the combined data with Delta V and Age as the independent variables are as follows:

Single-Vehicle Accident

(3.98) $\hat{p}_i = F(2.1130 - 0.0518X_1 - 0.0160X_2)$

CIA-2VEH

(3.95) $\hat{p}_i = F(3.5174 - 0.0912X_1 - 0.0177X_2)$

OID-2VEH

(3.96) $\hat{p}_i = F(3.8541 - 0.1090X_1 - 0.0136X_2)$

OIP-2VEH

(3.97) $\hat{p}_i = F(2.9543 - 0.0725X_1 - 0.0197X_2)$

where
$\hat{p}_i$ is the estimated probability of a non-severe injury,
F is the Logistic distribution function,
X_1 is Delta V, and
X_2 is Age.

Equations 3-98, 3-95, 3-96 and 3-97 state that the estimated probability of a non-severe injury, $\hat{p}_i$, is a function of Delta V and Age above. $\hat{p}_i$ is bounded by a value of zero and one. A $\hat{p}_i$ will certainly be severe and a value of one predicts a certainty of a non-severe injury. A $\hat{p}_i$ value greater than 0.5 will predict a non-severe injury while that less than 0.5 a severe injury.

The interpretation of these models may be simpler if Equations 3-98, 3-95, 3-96 and 3-97 are restated in terms of the estimated probability of a severe injury, which is simply $1-\hat{p}_i$. The models imply that:

1. The estimated probability of a severe injury ($1-\hat{p}_i$) increases as Delta V increases. A change of a Delta V value from 15 mph to 30 mph for a 30 year old occupant in the CIA-2VEH subset will result in the probability of a severe injury increasing from 0.05 to 0.39.

2. The estimated probability of a severe injury ($1-\hat{p}_i$) increases with Age. For the CIA-2VEH subset with Delta V of 30 mph, a change in Age from 30 to 60 will increase this probability from 0.39 to 0.62.

The models as represented by Equations 3-95 to 3-98 predicted overall injuries quite well. A closer examination of the prediction results revealed that if an injury was observed to be non-severe the models would be correct in the prediction well over 95% of the time, but if an injury was reported severe the models would only be correct in the prediction about 20% to 40% of the time. Although the number of non-severe injuries in front impacts outnumbered severe injuries by a ratio of 4 to 1 or more, it is important to be able to predict severe injuries reasonably well. Most, if not all, of the severe injuries which were mispredicted were those with low to medium Delta V values. These cases represented about 10% of total injuries and their severities depend, among other things, on the body regions being injured and what these body regions came into contact with at the impact.

In order to improve the predictive capability of the models, particularly in predicting severe injuries, many other variables were investigated with a view to incorporating them into the models. Some of these variables were Restraint Usage, Ejection, Body Region, Injury Type and Contact Point. It was found that over 95% of the occupants did not use or have restraints. This coupled with the likelihood that Restraint Usage were somewhat correlated with Age and/or Delta V made this variable of little value in the modelling. Ejection was found to be a slightly better variable than Restraint Usage. Over 75% of the occupants were not ejected or trapped. Ejection was found to be significant only when both the dummy terms and the interaction terms with Delta V were included in the model and only when the sample size was large. Ejection improved the predictive capability of the single-vehicle subset of the combined Phases 1 and 2 data by increasing the percent correct prediction of severe injuries by 7%.

The majority of the mispredictions were found to be associated with injury types such as rupture, dislocation, fracture, and hemorrhage, or with body regions such as abdomen, chest, pelvic/hip, and the lower limbs. These mispredicted cases were also found to be associated with the accidents where the occupants had come into contact with the front panels and/or the steering assembly of the vehicles. A couple of notable exceptions were severe ankle/foot injuries (fractures and

dislocations) and severe neck fractures. With the former, the contact points usually were floor and/or foot controls, while the latter did not appear to be caused by contact with any particular item.

It was thought that with Delta V, Age, Contact Point and Body Region in the models, the occurrence of injury severity would become more explainable. Delta V is a proxy for the force exerted on the occupant at the impact, causing the occupant to move from his/her original position and causes a body region to come into contact with the interior of the vehicle. The occupant's age could give rise to the resistance to or the tendency for a certain injury type.

However, the contact point variable did impose some serious problems in modelling. First, the missing data on Contact Point for both severe and non-severe injuries was to the extent that the number of valid cases for modelling would drop by about 50% for severe injuries and over 50% for non-severe injuries. Second, Contact Point showed a correlation with Body Region. Between Contact Point and Body Region, the latter appeared to be a better independent variable because a body region, with the exception of Head and Face, tended to be associated with either one or at most two major contact points. A contact point, on the other hand, could imply many different body regions.

That a body region almost immediately implies a certain contact point justifies its inclusion in the model even without another variable such as Contact Point. In fact, the final model for all subsets with Delta V, Age, and Body Region as the independent variables showed most considerable improvement in the model's predictive capability in that the percent correct prediction of severe injuries had considerably increased. The influence of the Body Region variable on the injury prediction can be seen from the following numerical example. In a two-car frontal accident with Delta V of 20 mph, an abdominal rupture to a 30 year old occupant (or the CIA-2VEH) will be predicted to have the probability of a severe injury of 0.11 (i.e., a non-severe injury) by the two-variable model (Delta V and Age). This same injury to the same occupant will be predicted by the model with Delta V, Age and Body Region to have the probability of a severe injury of 0.54 (i.e., a severe injury).

3.6 Summary

This section presented statistical models describing the relationship of injury severity to variables describing the occupant and the type and severity of the impact. In Section 2, these models were described as "mechanistic." The selection of independent variables and the form of the model were guided, to the degree possible, by consideration of the physical principles and mechanisms involved. The objective is to develop insight as to the significant variables influencing the probability of injury and their effects. Ultimately, such models are expected to be useful for evaluating modifications intended to improve vehicle crashworthiness.

This summary is organized under two headings: Analysis Techniques and Results. The salient procedures and methods used are briefly reviewed in the first subsection, while the second summarizes the results.

3.6.1 Analysis Techniques. The specific model used in this analysis is the logit model. The dependent variable, Injury Severity, is categorical. The logit model postulates that the probability of observing a particular category varies systematically with other variables that can be observed. For this analysis, the dependent variable, Injury Severity, was dichotomized into non-severe and severe.

The initial step in the model development was a subsetting of the data set with the objective of forming groups of occupants with similar injury production mechanisms. Single-vehicle accidents were separated from two-vehicle accidents. Vehicles in frontal impacts were separated from those in side impacts (other impact types were not studied). Occupants were separated based on seat location. In addition, occupant seat locations were sometimes identified as to whether they were "near" the impact point or "far" from the impact point. For example, in side impacts "near-side" occupants are separated from "far-side" occupants. The initial model development was carried out within the subsets.

A very important part of the model development process is the evaluation of the model. The Likelihood Ratio Statistic was used to test the significance of any particular model or to compare two models.

This statistic is influenced by sample size, so that it is somewhat difficult to compare results from subsets with different sample sizes. To complement this test, a goodness of fit was also computed. Having estimated the parameters for a particular model, a predicted probability of observing a non-severe injury can be computed for each case (occupant). The prediction is taken to be "correct" if the predicted probability is greater than 0.5 and the observed injury is, in fact non-severe, and so on. In this way, the number of correct and incorrect predictions is determined, and the goodness of fit is then measured as the percentage of correct predictions. This procedure identifies mispredictions which may be viewed as being similar to the residuals in a regression analysis. Studying the mispredictions is then a valuable tool in model development. Histograms of the predicted probabilities were presented for the non-severe and severe injuries separately. These histograms were central to the analysis procedure because they revealed that the models did not predict the occurrence of the severe injuries nearly as well as the occurrence of non-severe injuries. Models having an overall percentage of correct prediction of 80% or better were found to predict only 30% or so of the severe injuries correctly. Similarly, addition of a variable to the model would sometimes result in a significant Likelihood Ratio Statistic without improving the predictive capability appreciably. Model evaluation became the critical step in the model development procedure.

In the later stages of the model development, "outliers" were identified as the cases predicted incorrectly by the current model. These outliers were divided into non-severe injury cases incorrectly predicted and severe injury cases incorrectly predicted. Candidate variables for inclusion in the model were then studied in terms of their relationship to the outliers. For example, for a categorical variable, a two-way table of the variable levels versus the correct/incorrect prediction variable would be prepared in order to determine if certain levels of the candidate variable were correlated with the cases currently predicted incorrectly. This strategy frequently guided the subsequent grouping of the levels of a candidate variable before inclusion in the model. This procedure was used, in particular, for the

body region, injury type; contact point, restraint usage and ejection variables.

The final phase in the model development procedure was the plotting of the estimated logistic curves and their confidence intervals. Because Delta V was the dominant variable in these models, the predicted probability of a severe injury was plotted versus Delta V. Confidence limits on the predicted probability as a function of Delta V were also computed and plotted. The expressions for the variance were developed from a Taylor Series expansion and reflect the variance at a particular value of each of the independent variables, rather than the simultaneous confidence intervals which may be computed for a regression analysis. Comparing the logit curves for different models, or for different values of the independent variables in a particular model, is a convenient way to assess the differences between models or the magnitude of the effect of the independent variables. Such curves were produced in the final phase of the model evaluation process.

This subsection has summarized the analysis techniques used and, in particular, the model evaluation procedures. Model evaluation was the critical phase of the model development procedure. The next subsection summarizes the results of the development of mechanistic models.

3.6.2 Results. As expected, Delta V is the dominant variable in all of the models developed; the side impact models use the lateral component of Delta V. Age is the second variable which is common to nearly all of the models. In general, the Delta V and Age models correctly predict the injury severity 80-90% of the time.

A very important finding is that the Phase 1 models predicted injury in the Phase 2 data virtually as well as in the data set used to estimate the coefficients (the Phase 1 data). Even though Phase 2 is just a continuation of the original study, this stability is very reassuring. NCSS is the first study of sufficient scope to provide the opportunity for this kind of observation. Consequently, statistical tests indicated that it was appropriate to combine the Phase 1 and Phase 2 data. The resulting increase in sample size enhanced our ability to evaluate candidate variables.

The major problem in the model development was the prediction of severe injuries. Although the models developed in the early stages predicted injury quite well overall, the prediction of severe injuries was correct less than 40% of the time. Credit for the ability to make this distinction belongs with the careful model evaluation procedures used. The major thrust in the model development was the improvement of the prediction of severe injuries.

Body Region (of the injury) and Injury Type were found to be highly correlated with the misprediction of severe injuries. Careful incorporation of these variables into the models improved the predictive capability for severe injuries by as much as 30%. However, there are conceptual problems with including these variables. The Abbreviated Injury Scale is such that many injury types can only be assigned to one or possibly two AIS levels. In turn, particular injury types are associated with particular body regions; concussion only occurs in the head, and fractures tend to occur in the extremities. Consequently, specifying the body region and/or injury type comes close in many instances to specifying the AIS level. This is cheating. From the practical standpoint, inclusion of body region or injury type in the model is not useful without knowing the factors that determine which body region is injured, or what the type of injury will be.

Variables such as Principal Direction of Force and Contact Point were candidates to provide this kind of link. Attempts to incorporate these variables were fraught with problems. The first simply did not help the prediction at all. Missing data on Contact Point is over 50%. In addition, Contact Point tended to be associated with Body Region. However, the contact point did not discriminate as well as the body region. A given contact point was associated with several body regions. However, a given body region tended to be associated with only one, or possibly two, contact points. The association between the two variables was such that attempts to include both in the models produced anomalous results. Body Region also has less missing data than Contact Point. Consequently, the final models include the Body Region variable. The interpretation, or justification, for inclusion of this variable is that

it reflects both the contact point and the body region in a more effective manner than the current Contact Point allows on its own.

Statistical tests indicated that some of the original subsets could be combined without loss of significance. The final models were developed for two subsets of the side impacts and four subsets of the frontal impacts. These subsets are listed below:

Side Impacts

1. Impacts involving damage to the passenger compartment and occupants on the same side of the vehicle as the impact.
2. All far-side occupants (seated opposite the impacted side) and near-side occupants of impacts not damaging the passenger compartment.

Front Impacts

1. Single-vehicle accidents
2. Two-vehicle accidents with center impacts
3. Two-vehicle accidents with off-center impacts, drivers only
4. Two-vehicle accidents with off-center impacts, right front passengers

The models for side impacts which have Lateral Delta V and Age as the independent variables are shown in Section 3.3.3 (Equations 3-57 and 3-58) and that goodness of fit it Table 3.24. The side-impact models which include two additional variables, Object Contacted and Body Region are shown in Section 3.3.4 (Equations 3-60 and 3-61) and their goodness of fit results are shown in Table 3.31.

The models for front impacts which have Delta V and Age as the independent variables are shown in Section 3.5.5 (Equations 3-95 to 3-98) and their goodness of fit results are shown in Table 3.62. The front-impact model with Delta V, Age and Body Region are shown in Section 3.5.7 (Equations 3-100 to 3-103) and their goodness of fit in Table 3.65.

Several dummy variables were used to bring in the various categories of Body Region. These models predicted non-severe injuries correctly about 90% of the time and severe injuries correctly from 43% to 67% of the time.

An alternative approach which now appears attractive is to develop separate models for each Body Region. In effect, one would be studying the crashworthiness of vehicles separately for heads, chests,

extremities, etc. This approach requires that injury information be collected for each body region of interest. In particular, the occurrence of no injury to a particular body region would have to be recorded. Since all injuries were not recorded in the NCSS study, only the first six, one cannot be sure that a given body region was not injured just because none is recorded. However, making such an assumption may not be too unreasonable. This approach would seem to merit further work.

A final area worth noting concerns the use of that Abbreviated Injury Scale (and even a simple dichotomy of this variable) as the dependent variable. Virtually all of the outliers (mispredictions) were relatively low collision severity (Delta V) impacts which resulted in severe injuries (OAIS 3+). A case-by-case examination of these indicated that the problem seems to lie in the quantification of injury. Many factors were assimilated in developing the AIS Scale: threat-to-life, treatment period, probability of permanent impairment, etc. Not all of these factors are directly related to the collision forces. For example, a substantial group of outliers in the frontal models were ankle/foot dislocations or fractures. These injuries receive an AIS 3. However, the forces which produce these injuries would not seem to be particularly high. Often it would seem that a simple broken bone or contusion might result. The critical factor is probably the specific point and direction of force application. This kind of problem will hamper further development of mechanistic models.

There are probably several approaches to this problem. Developing separate models for each Body Region will somewhat extend the usefulness of the AIS scale, since particular injury types are usually associated with a given body region. However, the number of possible AIS levels is usually reduced to 2 or 3 in this situation. A more difficult alternative is to continue the original philosophy advanced by the the developers of the AIS scale[23]; that is, to develop several scales, each addressing the different dimensions of injury severity. In this way injuries would be rated on separate scales for threat-to-life, treatment

---

[23] The Abbreviated Injury Scale (1976 revision). (Morton Grove, Ill., American Association for Automotive Medicine, 1976).

period (or even cost), likelihood of permanent impairment, energy required to produce the injury, etc. A multi-dimensional dependent variable would then be available.

4 POPULATION STATISTICS

One of the major uses of the NCSS data is to provide nationally representative accident statistics for the U.S. population. This subsection describes the analyses used to accomplish this task. The NCSS data provides estimates for the aggregate of the seven areas chosen for NCSS. The methodology used to produce these estimates is discussed in Section 4.1. These statistics were organized for publication, and the books of NCSS statistics produced are discussed in Section 4.3. The sampling errors for selected statistics were calculated and these results are presented in Section 4.4.

Missing data will affect all of the statistics calculated using the NCSS data. The two major sources of missing data are missing accidents in the collection of the data and incomplete information on those accidents collected. In Section 4.2 possible sources for missing accidents are discussed and the effect on the NCSS statistics is evaluated. The effect of missing data on OAIS and Delta V distributions is discussed in Section 4.5.

A method of modifying the NCSS data to provide nationally representative estimates is used in Section 4.6 and some selected national projections are presented. In Section 4.2 the areas included in NCSS are compared to the nation in an attempt to ascertain the representativeness of the NCSS areas.

This section is organized in the following manner. Section 4.1 presents all theoretical foundations for work done in subsequent subsections. Sections 4.2 to 4.6 present the analytic work done in each of the areas described in Section 4.1. Section 4.7 provides a summary of the main results from the analysis described in this section.

4.1 Analytical Technique - Weighted Analysis

This subsection is intended to document estimation procedures and sampling error procedures used in the analysis of the NCSS data. Also included here is a discussion of alternative approaches for presenting sampling errors for large scale surveys. Procedures for missing data adjustments are reviewed and the rationale for the chosen missing data analysis is presented. Finally the issue of estimating "nationally

representative" statistics is discussed and a procedure developed to produce national projections.

4.1.1 Sample Design. The National Crash Severity Study was a data collection effort oriented towards collecting data to relate injury severity to crash severity. The origin of the NCSS design is described in a paper by Kahane, Smith, and Tharp.[24] The study design consisted of the purposive selection of seven areas, chosen to represent the United States population distribution. These seven areas include 41 counties, 2 partial counties, and 3 police districts in the city of Los Angeles.

Within each area a sample design was specified for the sampling of accidents for that area. An accident was eligible for sampling if it included a towed passenger car (this definition was extended in Phase 2 of the study to include passenger cars, light trucks and vans) and if the most severe injury to any occupant in the accident was to an occupant of a towed vehicle.

There were two basic sample designs used. In two of the seven areas accidents were selected with probabilities related to the severity of injury in the accident. Those accidents that involved a fatality or someone who was transported to the hospital and kept overnight were selected with certainty (Stratum 1). Accidents that had at least one occupant of a towed vehicle transported to the hospital but not kept were selected at a rate of 25 percent (Stratum 2). Those accidents where none of the occupants were transported to the hospital (those with minor injuries and those with no injury) were selected at a rate of 10 percent (Stratum 3). This sample design was used by the Highway Safety Research Institute and Southwest Research Institute.

Accidents in the other five areas were sampled using two independent systematic samples of days to sample accidents that did not involve transportation to a hospital (Stratum 3) and accidents that required transportation to the hospital but not hospitalization (Stratum

---

[24] C. J. Kahane, R. A. Smith, and K. J. Tharpe (sic), "The National Crash Severity Study," International Technical Conference on Experimental Safety Vehicles. Sixth Report (Washington, D.C.: National Highway Traffic Safety Administration, 1978), pp. 493-516.

2). Within each of the areas every accident involving a fatality or an occupant transported and kept overnight at the hospital was investigated (Stratum 1). A systematic sample of days choosing every tenth day starting with a randomly chosen day was used in Stratum 2 and all accidents involving minor or no injury to an occupant that occurred on that day were sampled. An exception was Los Angeles which used a systematic sample choosing every five days and gave each accident with minor injury that occurred on that day an equal probability of being included or excluded in the study. A systematic sample of days choosing every fourth day starting with a randomly chosen day was used to sample accidents in Stratum 3--those involving an occupant who was transported to the hospital and released. Again every accident of this type that occurred on the sampled day was investigated.

NCSS data was collected in two phases. Phase 1, the first fifteen months, included in the population only passenger cars that were towed from the scene of the accident for damage. Phase 2 of the study involved data collection for the following twelve months. In this phase the population was extended to include all light trucks and vans that were towed from scene of the accident for damage. In Phase 2, Southwest Research Institute, which had been sampling accidents with probabilities related to the severity of injury in the accident, changed the rate of sampling accidents that involved minor injuries in Stratum 3. These accidents were sampled at a rate of 5 percent.

It was not always the case that the same systematic sample was used for each stratum within a particular study area. Some teams chose different starting days for their systematic sample of days depending on county. A term was needed to identify within each study area the group of counties in which the sampling procedure was identical. The term we have chosen is "design group." A summary of the sample designs for each design group within each stratum is given in Table 4.1. This table defines 10 design groups.

4.1.2 Estimation Methodology. The sample designs described in the preceding subsection allow the development of estimates of statistics related to accidents within each of the areas chosen. These estimates have certain statistical properties characteristic of good estimates.

TABLE 4.1

Description of Sample Designs
within the NCSS Sites

Design Group	First Stage Sampling Unit	Stratum 3	Stratum 2
Calspan	Days	10% Systematic Sample starting Jan. 7, 1977	25% Systematic Sample starting Jan. 1, 1977
Highway Safety Research Inst.	Accidents	Equal probability of selection (p=.10)	Equal probability of selection (p=.25)
Southwest Research Inst.	Accidents	January 77 - March 78 Equal probability of selection (p=.10) April 78 - March 79 Equal probability of selection (p=.05)	Equal probability of selection (p=.25)
Miami	Days	10% Systematic Sample starting Jan. 3, 1977	25% Systematic Sample starting Jan. 4, 1977
Indiana[a] Group A	Days	10% Systematic Sample starting Jan. 3, 1977	25% Systematic Sample starting Jan. 3, 1977
Indiana[b] Group B	Days	10% Systematic Sample starting Jan. 4, 1977	25% Systematic Sample starting Jan. 4, 1977

[a]Includes counties of Bartholomew, Brown, Daviess, Dubois, Gibson, Lawrence, Martin, Monroe, and Pike.

[b]Includes counties of Greene, Jackson, Knox, Owen, Perry, Posey, Spencer, and Warrick.

TABLE 4.1 (Continued)

Design Group	First Stage Sampling Unit	Stratum 3	Stratum 2
Los Angeles	Days	20% Systematic Sample starting Jan. 2, 1977 with an equal probability sample of accidents on a chosen day (p=.50)	25% Systematic Sample starting Jan. 4, 1977
Kentucky[a] Group A	Days	10% Systematic Sample starting Jan. 1, 1977	25% Systematic Sample starting Jan. 1, 1977
Kentucky[b] Group B	Days	10% Systematic Sample starting Jan. 2, 1977	25% Systematic Sample starting Jan. 2, 1977
Kentucky[c] Group C	Days	10% Systematic Sample starting Jan. 3, 1977	25% Systematic Sample starting Jan. 3, 1977

[a]Includes counties of Clark, Jessamine, and Madison.

[b]Includes counties of Woodford, Scott, and Bourbon.

[c]Includes Fayette county.

Since the areas were chosen purposively these accident statistics generalize only to the specific areas chosen. Extrapolation of estimates for areas observed to areas unobserved, generating a national projection that describes the national accident experience, will be presented and discussed in Section 4.6.

Of particular interest is the estimation of proportions of certain categories for variables that are collected in NCSS. Estimates of this type were used to produce the various statistics for the publications describing the NCSS statistics. These publications will be discussed in Section 4.3. The estimates can also be used to develop models, categorical models or linear models, that describe relationships between the estimates of the population totals or proportions. These models are described in Section 4.7. For each of these tasks the estimation procedure for means and proportions is the same. The estimation procedure used to estimate means and proportions for accident, vehicle, and occupant level statistics is presented and discussed below.

All of the sample designs described in Table 4.1 involve cluster sampling. A cluster is defined by either an accident (for Highway Safety Research Institute and Southwest Research Institute) or a day (for all other teams). Within each accident or day selected all vehicles that were towed were chosen for further investigation (except for Los Angeles where only half of the accidents involving a towed vehicle were selected in Stratum 3). Within each chosen vehicle all occupants in that vehicle were investigated. So there are three basic sets of variables; the accident variables, the vehicle variables and the occupant variables. The accident variables describe the accident environment at the time of the crash. The vehicle variables describe the vehicle's status at the time of the accident and also include information about the damage induced by the collision. The occupant variables describe the occupant and the resulting injuries the occupant sustained in the accident. So for each cluster a total or a count can be obtained for all accidents, vehicles or occupants represented in the cluster. When sampling clusters, the clusters become the basic unit of observation and cluster totals are used as a basis for estimates and are used in the calculation of the variance of these estimates.

In order to estimate means and proportions it is necessary to view these statistics as a ratio of two sums. The sum in the numerator is the estimate of the total for a variable or the count of the number of items in a particular category. The sum in the denominator is the estimate of the number of items used in the calculation of the numerator statistics. More specifically, let $\hat{M}$ be an estimate of a population mean involving accidents (or vehicles or occupants). The mean of the variable, X, to be estimated is given by

$$\hat{M} = \hat{X}/\hat{N} \tag{4-1}$$

where

$$\hat{X} = \sum_{i=1}^{3}\sum_{k=1}^{m_i} x_{ik}/p_i, \tag{4-2}$$

$$\hat{N} = \sum_{i=1}^{3}\sum_{k=1}^{m_i} n_{ik}/p_i, \tag{4-3}$$

and

x_{ik} is the variable total for the accidents (or vehicles or occupants) in the k^{th} cluster in the i^{th} stratum, (i=1,2,3),

n_{ik} is the number of accidents (or vehicles or occupants) in the k^{th} cluster in the i^{th} stratum,

p_i is the probability of selection in the i^{th} stratum, and

m_i is the number of clusters in the i_{th} stratum.

The estimate of a population proportion[25] of Category A, $\hat{p}(A)$, is given by,

$$\hat{p} = \hat{T}/\hat{N} \tag{4-4}$$

where

$$\hat{T} = \sum_{i=1}^{3}\sum_{k=1}^{m_i} t_{ik}/p_i, \tag{4-5}$$

$$\hat{N} = \sum_{i=1}^{3}\sum_{k=1}^{m_i} n_{ik}/p_i, \tag{4-6}$$

[25] In the following notation reference to category A will be suppressed, the estimate of P(A) will be written $\hat{p}$. $\hat{p}$ is then the estimated proportion of category A.

and

t_{ik} is the number of accidents (or vehicles or occupants) of Type A in the k^{th} cluster in the i^{th} stratum, (i=1,2,3)

n_{ik} is the number of accidents (or vehicles or occupants) in the k^{th} cluster in the i^{th} stratum,

p_i is the probability of selection in the i^{th} stratum, and

m_i is the number of clusters in the i^{th} stratum.

It should be noted that in calculating accident level statistics for a sample of accidents each cluster represents only one accident. So, in Equation 4-3 and Equation 4-6, $\hat{N}$ simplifies to

$$(4\text{-}7) \qquad \hat{N} = \sum_{i=1}^{3} m_i/p_i$$

where p_i and m_i are as defined for Equations 4-1 to 4-3.

For those teams using a systematic sample of days, this estimation process ignores the additional fact that the selected clusters were obtained systematically. Estimation procedures specifically derived for systematic sample designs involve intracorrelations. Information about the magnitude of these intracorrelations is not known. Therefore as an approximation the systematic sample of days was treated as a simple random sample of days.

Under the assumption of simple random sampling, $\hat{M}$ defined by Equation 4-1 and $\hat{p}$ defined by Equation 4-4 are approximately unbiased estimates of the true mean or proportion. Since the mean and the proportion are ratios of two variables, the expected value calculated for the ratio is the expected value of the Taylor Series approximation to the ratio. This method of approximation is discussed in Cochran.[26] It implies that, on the average, over repeated independent samples, the ratio estimate will be a good approximation to the true population mean or proportion. For cluster designs there is no exactly unbiased estimate of the population mean or proportion.

---

[26] William G. Cochran, Sampling Techniques, 3rd ed. (New York: John Wiley & Sons, 1977), p. 319.

4.1.3 Estimation of Variance. For every estimate an associated variance or standard error can be estimated. These variance estimates depend on the particular design used to select the primary units. In general, the variances decrease as the number of primary units increase. Since the NCSS design has well defined sampling designs within each area, it is possible to calculate an estimated variance for estimates of means and proportions for a given area. Since areas were chosen independently, these individual variance estimates can be used to obtain an overall estimate of variance for the aggregate of the areas.

It is important to calculate the sampling errors associated with the statistics calculated. These sampling errors give a direct measure for the reliability of the estimate calculated. If the standard error is high, then a fair amount of uncertainty can be associated with the estimate. On the other hand, if the variance is quite small compared with the magnitude of the estimate, then the reliability of the estimate is enhanced.

In this Section attention will center on estimated variances for proportions estimated. The formulae used are exactly the same if the estimated variance of a mean is required.

The estimation of the variance of a proportion is specific to the design of the sample. In NCSS, accident level statistics for HSRI and SwRI are calculated differently than vehicle or occupant level statistics. The variance for the latter calculation is exactly the same as the variance calculation for those teams sampling days. Therefore the exception will be discussed first and following that the more general calculation of variance.

When the cluster in the sample design is accidents and concern centers on accident statistics, the estimate is given by Equation 4-4. As noted in the previous discussion in this special situation there is only one accident per cluster. This implies that there is no variability in the number of accidents in a cluster. In doing the actual sampling the total number of accidents was not recorded so that the estimated number of accidents given by Equation 4-7 must be assumed to be known without variability. Rewrite the numerator and denominator of the proportion, Equations 4-5 and 4-6, as

$$(4\text{-}8) \qquad \hat{T} = \sum_{i=1}^{3} \hat{t}_i / p_i,$$

$$(4\text{-}9) \qquad N = \sum_{i=1}^{3} n_i / p_i,$$

where

$\hat{t}_i$ is the total number of accidents of Type A in the i^{th} stratum,

n_i is the total number of accidents in the i^{th} stratum and

p_i is the probability of selection of clusters in the i^{th} stratum.

The variance of the ratio of $\hat{T}$ to N is given by

$$(4\text{-}10) \qquad \text{Var } \hat{T}/N = [\text{Var } \hat{T}]/N^2$$

since the denominator, N, is assumed to be known. The variance of the numerator is given by

$$(4\text{-}11) \qquad \text{Var } \hat{T} = [\text{Var } \hat{t}_2]/p_2^2 + [\text{Var } \hat{t}_3]/p_3^2$$

since $\hat{t}_1$ has no variance (sampled at 100%).

In order to estimate this variance an unbiased estimate (or approximately unbiased estimate) is used in Equation 4-11 and the estimated variance is given by

$$(4\text{-}12) \qquad \hat{\text{Var}}\ \hat{T} = [\hat{\text{Var}}\ \hat{t}_2]/p_2^2 + [\hat{\text{Var}}\ \hat{t}_3]/p_3^2$$

where $\hat{\text{Var}}\ \hat{t}_2$ and $\hat{\text{Var}}\ \hat{t}_3$ are the simple estimated variances of the cluster totals within strata defined in Equation 4-15. Variance calculations for means are completely analogous.

In all other situations, estimated means and proportions are ratios of two random variables, the numerator sum and the denominator sum. The estimated variance of the ratio of two random variables is customarily approximated by

$$(4\text{-}13) \qquad \hat{\text{Var}}\ \frac{\hat{T}}{\hat{N}} \doteq \hat{p}^2 \left[\frac{\hat{\text{Var}}\ \hat{T}}{\hat{T}^2} + \frac{\hat{\text{Var}}\ \hat{N}}{\hat{N}^2} - 2\,\frac{\hat{\text{Cov}}(\hat{T},\hat{N})}{\hat{T}\hat{N}} \right]$$

The estimated variances in Equation 4-13 are given by:

$$(4\text{-}14) \qquad \hat{Var}\ \hat{T} = [\hat{Var}\ \hat{t}_2]/p_2^2 + [\hat{Var}\ \hat{t}_3]/p_3^2$$

where

$$(4\text{-}15) \qquad \hat{Var}\ \hat{t}_i = \sum_{k=1}^{m_i} (t_{ik} - \bar{t}_i)^2/(m_i-1) \qquad i=2,3$$

and

$$\bar{t}_i = \sum_{k=1}^{m_i} t_{ik}/m_i \qquad i=2,3;$$

$$(4\text{-}16) \qquad \hat{Var}\ \hat{N} = [\hat{Var}\ \hat{n}_2]/P_2^2 + [\hat{Var}\ \hat{n}_3]/P_3^2$$

where

$$(4\text{-}17) \qquad \hat{Var}\ \hat{n}_i = \sum_{k=1}^{m_i} (n_{ik} - \bar{n}_i)^2/(m_i-1) \qquad i=2,3$$

and

$$(4\text{-}18) \qquad \bar{n}_i = \sum_{k=1}^{m_i} n_{ik}/m_i \qquad i=2,3;$$

finally

$$(4\text{-}19) \qquad \hat{Cov}\ [\hat{T},\hat{N}] = \hat{Cov}[\hat{t}_2,\hat{n}_2]/p_2^2 + \hat{Cov}[\hat{t}_3,\hat{n}_3]/p_3^2$$

where

$$(4\text{-}20) \qquad \hat{Cov}[\hat{t}_i,\hat{n}_i] = \sum_{k=1}^{m_i} (\hat{t}_{ik} - \bar{t}_i)(\hat{n}_{ik} - \bar{n}_i)/(m_i-1)$$

Using Equation 4-13 an approximation to the variance of the estimate can be obtained. Variance calculations for the mean are done in a similar manner.

The finite population correction factor has been eliminated from the variance formulae discussed above. The effect of this is to slightly over-estimate the variances if the model of simple random sampling is the appropriate model. The sample design was not based on a simple random sample of days. Therefore these formulae are approximations and omitting the finite population correction factor seemed justified.

Treatment of sampling errors for a large-scale survey where many different variables are collected can be difficult, both in the calculation and presentation. Currently there are two ways of presenting sampling errors. One method graphically presents estimators and their variances. So from the calculated value of the estimate alone an approximation of the sampling error can be interpolated from the graph.

Another method of summarizing the effect of the design on the variance of an estimate is given by the quantity called the design effect. The design effect is discussed by Kish[27] and is defined by

$$\text{(4-21)} \qquad \text{Design Effect} = \hat{\text{Var}}\,(\hat{X}|\text{Design})/\hat{\text{Var}}\,(\hat{X}|\text{SRS})$$

where

$\hat{\text{Var}}\,(\hat{X}|\text{Design})$ is the variance of $\hat{X}$ under the given design and

$\hat{\text{Var}}\,(\hat{X}|\text{SRS})$ is the variance of $\hat{X}$ if simple random sampling is assumed, and $\hat{X}$ is the estimated mean or proportion.

In general, for cluster sampling, variables will have design effects greater than one. Design effects for variables obtained from a stratified simple random sample will generally be less than one. When stratification and cluster sampling are used together the design effects will depend on how stratification and clustering interact for the particular variable being estimated.

4.1.4 Missing Data Adjustments. In most large-scale data collection projects the problem of missing data needs to be seriously

---

[27] Leslie Kish, Survey Sampling (New York: John Wiley & Sons, 1965), pp. 257-259.

addressed. The problem of nonresponse is complex because of the many kinds of nonresponse possible. This discussion will focus on two particular types of nonresponse, unit and item nonresponse. Unit nonresponse occurs when no information can be obtained about the chosen accident, vehicle, or occupant. The only information available is that data should have been obtained but was not collected. Item nonresponse involves nonresponse to a few of the variables collected but full information on the remaining variables. So there is always some information for the particular case. A third type of non response, considered in Section 4.2, is undercoverage of accidents.

There are two major types of adjustments for nonresponse. The first type involves a reweighting of the data to adjust for the missing data. These procedures are generally useful for unit nonresponse problems. Alternatively there are imputation techniques that impute for the missing data a "best" predicted value. Chapman[28] summarizes most of these techniques and they will be briefly outlined here.

Reweighting procedures involve a reweighting of the means of different classes within the data set. Stratification of the data with a reweighting inversely proportional to the response rate within each stratum is an example of this type called "weighting class adjustments." Here the stratification is chosen such that the strata are homogeneous and have different response rates within each stratum. Raking ratio estimates involve using external data and iterative proportional fitting to adjust for missing data. Double sampling procedures view the population as two populations, the responding population and the nonresponding population. The first sample gives information about the responding population and the nonresponding population is again sampled to try to obtain an estimate for the nonresponding population. The resulting estimate for the sample is a weighted average of these two means.

---

[28] David W. Chapman, "A Survey of Nonresponse Imputation Procedures," American Statistical Association Proceedings of the Social Statistics Section, 1976: Part I (Washington, D.C.: American Statistical Association, 1976), pp. 245-251.

Imputation procedures, on the other hand, are techniques to substitute a value for a missing data item. The "hot deck" procedure[29] is used widely. Cells are defined by variables with no missing data to form homogeneous groups. Then cases with a missing item draw their value from one of the cases in the cell in which the case falls. How the case is chosen determines variations on this procedure. Statistical matching can also be used to find the "nearest neighbor" and the value is then imputed based on the match. The EM algorithm[30] is another method of imputation that imputes a value for a missing data item.

All of the procedures currently used make an assumption about the nonresponding units. Technically, the items need to be "missing at random."[31] What this means is that the nonresponding population is not "special" in any way. To be missing at random does not preclude a different distribution of cases across a certain variable but does exclude the case where all extreme values fall into the nonresponding population. The basic assumption is that the nonrespondents must be like the respondents.

A double sampling approach was chosen in this investigation of the NCSS missing data. The assumption of "missing at random" is a crucial assumption in all imputation procedures and a double sample allows an investigation into possible deviations from this assumption. This approach allows for estimation of the sample distributions in the missing and non-missing subpopulations and makes possible the comparison of these two sample distributions. This approach does not lead to aggregate estimates for the NCSS population since the proportion of the aggregate that belongs to each subpopulation is not known. The result

---

[29] Innis G. Sande, "Hot Deck Imputation Procedures," Symposium on Incomplete Data: Preliminary Proceedings (Washington, D.C., Social Security Administration, December, 1979), pp. 484-507.

[30] A. P. Demster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," Journal of the Royal Statistical Society, ser. B, 39:1 (1977), 1-38.

[31] D. B. Rubin, "Multiple Imputations in Sample Surveys--A Phenomenological Bayesian Approach to Nonresponse," Imputation and Editing of Faulty or Missing Survey Data (Washington, D.C.: U. S. Department of Commerce, Bureau of the Census, 1978).

on this analysis is an estimate of the distribution of each subpopulation for the data sampled and an adjusted total for the data collected. The subsampling procedure is discussed in Section 4.5.1 and 4.5.3 and the results of the analyses can be found in Sections 4.5.2, 4.5.4, and 4.5.5.

4.1.5 Inference to the National Population. NCSS fails to satisfy the requirements of a national probability sample. The NCSS study design involved choosing seven areas in the United States to collect accident data for a study of the relationship of injury severity to crash severity. This choice depended on the availability and willingness of an investigating team to perform the necessary data collection. The areas covered by the teams were chosen in such a way that the percentage of residents in urban areas was approximately the same for the chosen areas as for the U.S. population as a whole. Within each area a sampling plan was used to select a sample of accidents. These sample designs varied from stratified random samples to cluster samples of days where all or some subsample of accidents were chosen.

Since the choice of the primary unit, the area, was not specified by a sample design, there is no single universally accepted method for producing national estimates from NCSS. The method developed for national estimates (which might more properly be called "national projections") uses the data collected in the chosen areas together with with demographic information available for all areas in the United States. The technique is based on the assumption that an observed relationship between accident statistics and demographic variables in the NCSS areas is appropriate for and adequately describes the relationship in the areas that have not been observed. The bias in the national projection is investigated, and an estimate of the variability of the national projection is produced.

The development of a national estimate from the NCSS design requires that certain assumptions be made concerning the population and the study design. The population is to be thought of in terms of a group of smaller subpopulations. For the United States these could be states, counties, or enumeration districts. Once the subpopulation is

specified, the total for the population can be defined in terms of the sum of the totals for each subpopulation.

The following procedure makes use of relationships between two groups of statistics collected for each specific subpopulation. In the first group, statistics may come from a complete census of each subpopulation or a physical description of the subpopulation. For example, FARS (Fatal Accident Reporting System) and the County and City Data Book[32] provide census-type statistics by county for all counties in the United States. It is essential for this procedure that there is at least one statistic that is known for every subpopulation in the population. These statistics form a group of statistics which have values for every subpopulation.

The second group of statistics that need to be expanded nationally, are known for only some of the subpopulations in the population. It is necessary that these statistics be unbiased estimates for their subpopulation totals and that an estimate of the variance of the total be obtainable using the sample design. NCSS provides accident statistics for 43 of the 3112 counties. These statistics are unbiased and a variance can be calculated for each accident statistic. This is possible because there is a specified sampling plan for NCSS data collection.

Finally it is assumed that any relationship between statistics in the two groups for the observed subpopulations will adequately describe the relationship in the entire population. More specifically, the prediction at a given value based on the observed data will be close to the prediction for that value if all the data were available.

Now, for the general framework, let $Y=(Y_1,\ldots,Y_N)$ be the population totals for the N subpopulations which together form the population. From an independent source $X=(X_1,\ldots X_N)$ is available and represents population totals on related variables of interest. X is known and available for every subpopulation. If the subpopulation totals, Y_i,

---

[32]U. S. Bureau of the Census, County and City Data Book, 1977: A Statistical Abstract Supplement (Washington, D.C.: Government Printing Office, 1978).

were known, investigation of the relationship between Y and X could be investigated directly and functions fit to the data to describe the relationship in the data for the population. With complete data about the target population, there may be many possible functions that might describe the relationship. The objective is to define a function that closely approximates the data and is still a simple function. Without complete data, past experience or some investigation will be necessary to specify a possible functional relationship. Once the function, $f(X)$, is specified every subpopulation can be described by the following:

$$(4\text{-}22) \qquad Y_i = f(X_i) + e_i$$

where $f(X_i)$ is the value of the function at X_i and e_i is the deviation of the subpopulation total from its functional value. Note that by specifying different functions it is possible to modify the magnitude of the e_i's.

Consider the following example. Let Y_i be the number of accidents in the i^{th} county and X_i be the retail gas sales in the same county. Suppose it is reasonable to assume that

$$(4\text{-}23) \qquad f(X_i;a,b) = a + b\,(X_i-\bar{X}) \qquad i=1,\ldots,N$$

where $\bar{X}$ is the population mean of X. That is, the number of accidents per county is a linear function of the retail gas sales of the county. The least squares estimates for a and b are given by,

$$(4\text{-}24) \qquad a = \sum_{i=1}^{N} Y_i/N$$

$$(4\text{-}25) \qquad b = \frac{\sum_{i=1}^{N} Y_i(X_i-\bar{X})}{\sum_{i=1}^{N} (X_i-\bar{X})^2}$$

Once a and b have been defined, the number of accidents in each sub population can be expressed as $Y_i = a + b(X_i-\bar{X}) + e_i$ where $a + b(X_i-\bar{X})$ represents that part of Y_i described by the function, and e_i measures the deviation from the regression line. With this definition of a and b

the sum of the e_i for the population is zero so that the population total is equal to the sum of $a + b(X_i - \bar{X})$ for all counties.

The model specified above describes all of the elements in a given population. If complete information on Y and X were known the entire population could be described and modelled with complete certainty. The problem of making national projections presents itself because complete data on Y is not available. The ultimate goal is to estimate a population total with the data available.

The population projection is defined to be an estimate of the expected value of the population total given the sample, $\theta(s)$. The expected value of the population total given the sample can be expressed as

$$(4\text{-}26) \qquad \theta(s) = E\left(\sum_{i=1}^{N} Y_i \mid \text{sample}\right)$$

$$= \sum_{i \in s} Y_i + \sum_{i \notin s} E(Y_i \mid \text{sample})$$

where the index of summation $i \in s$ means summation over all chosen counties and $i \notin s$ means summation over all counties not chosen. So that, theoretically, $\theta(s)$ is equal to the sum of the subpopulation totals for chosen counties plus the sum of the expected value given the sample (which may depend on the sample) for all counties not chosen.

The estimates of the population totals for the counties chosen are relatively straightforward. The assumption has been made that within each county there is an unbiased estimate for the county population total. Let $t=(t_1,\ldots,t_n)$, where n is the number of counties chosen, be the unbiased estimates for totals of the chosen counties. So that the sum of the t_i's for the observed counties is an unbiased estimate of the sum of the true populations for the observed counties.

A national projection can be obtained if an estimate of the population total for each county can be defined for those counties that were not observed. This estimate for the expected value of an unobserved county is a bit more arbitrary. Based on the population

model described above, if the function is chosen well, the e_i may be small enough so that $a + b(X_i - \bar{X})$ may be a reasonable approximation for Y_i.

The data from the counties observed is all the information available about the relationship implicit in the population. This information is used to estimate the expected value in an unobserved county. Consider, as an estimate of the expectation of the total of a county not observed, the estimate, $a + b(X_i - \bar{X})$, where a and b are chosen to minimize

$$(4\text{-}27) \qquad \sum_{i \epsilon s} [t_i - a - b(X_i - \bar{X})]^2$$

The estimates $\hat{a}$ and $\hat{b}$ are given by

$$(4\text{-}28) \qquad \hat{a} = \sum_{i \epsilon s} t_i / n$$

$$(4\text{-}29) \qquad \hat{b} = \frac{\sum_{i \epsilon s} t_i (X_i - \bar{x}_s)}{\sum_{i \epsilon s} (X_i - \bar{x}_s)^2}$$

where $\bar{x}_s$ is the mean of X for the units observed. In the previous example X would be retail gas sales. The national projection, in the linear case, will be defined by the following statistic:

$$(4\text{-}30) \qquad \hat{\theta}(s) = \sum_{i \epsilon s} t_i + \sum_{i \notin s} [\hat{a} + \hat{b}(X_i - \bar{x}_s)] .$$

Taking the expectation of $\hat{\theta}(s)$, the national projection, over all possible samples within each county for each county observed, it follows that,

$$(4\text{-}31) \qquad E\hat{\theta}(s) = \sum_{i=1}^{N} Y_i - \sum_{i \epsilon s} \varepsilon_i + \beta(N-n)(\bar{X} - \bar{x}_s) + \sum_{i \notin s} [(E\hat{\alpha} - \alpha) + (E\hat{\beta} - \beta)(X_i - \bar{x}_s)]$$

It can be seen from Equation 4-31 that the national projection is not unbiased. The bias, in the linear case, is a function of three quantities:

1. the difference between $a + b(X_i - \bar{x}_s)$, the prediction at X_i of a model based on the whole population, and $E\hat{a} + E\hat{b}(X_i - \bar{x}_s)$, the prediction at X_i of a model based on the expected value of the totals observed,

2. the difference between the mean value of X for the whole population and the mean value of X for observed counties, and

3. the deviations of the sampled counties from the true linear function.

One of the basic assumptions of the procedure is that the first quantity is zero. The other two terms that contribute to the bias are both functions of the particular counties observed. Since X is known for all counties, the difference $(\bar{X} - \bar{x}_s)$ can be calculated. Although b is still unknown, some information concerning this part of the bias can be obtained by using the estimate of b.

The bias due to the sum of the true deviations for the observed units is more difficult to assess. The sum of the theoretical deviations for the observed subpopulations may be large or small, positive or negative. The magnitude of this sum, if it is large, may be due to data which does not exhibit a strong relationship or it might result from an unfortunate choice of "deviant" units. There is no method at this time to assess the magnitude of this part of the bias. However if enough is known about the relationship and the population, it may be possible to minimize this term in the decision stage by the right choice of subpopulations.

The assumption has been made that the relationship between statistics in the chosen counties also holds for those counties not observed. Whether this is justified or not cannot be evaluated directly. An assessment of whether it is reasonable must be made for every national projection that is developed. Until there is evidence to support the position that the assumption is not valid it provides a useful framework to develop an estimate which may closely approximate the national total.

The variance of the national projection is given by the following formula:

(4-32)
$$V\hat{\theta}(s) = \frac{N^2}{n^2}\sum_{i\in s} c_i + \frac{\left[\sum_{i\in s}(X_i-\bar{X}_s)\right]^2}{\left[\sum_{i\in s}(X_i-\bar{X}_s)^2\right]}\sum_{i\in s}\left[\frac{(X_i-\bar{X}_s)^2}{\sum_{k\in s}(X_k-\bar{X}_s)^2}+\frac{2(X_i-\bar{X}_s)}{n}\right]Vt_i$$

where Vt_i is the variance of the estimate of the subpopulation total in the i^{th} chosen area. One method of obtaining an unbiased estimate of $V\hat{\theta}(s)$ is to substitute an unbiased estimate for the Vt_i. Since it was assumed that there was a sampling plan defined within each chosen county, an unbiased (or approximately unbiased) estimate can be obtained.

4.2 Sample Design Implications

In this subsection the relevance of the sample design is discussed in terms of its effect on the modelling effort and the estimation of accident statistics. Modelling requires a broad range of data so that models developed will be applicable fairly generally. Various proportions based on demographic variables for the NCSS area are compared with the national proportions. The characteristics of the areas with respect to urbanization and region of the country are examined.

The estimation of accident statistics requires a strict adherence to the sample design specified for accident collection. Only if all the data are obtained can it be expected that good reliable estimates will be produced. Here various sources of undercoverage are considered that may cause the NCSS statistics to be underestimates of the true value.

4.2.1 Sample Representativeness. A discussion of the background for the purposive selection of areas is given by Kahane, Smith, and Tharp[33]. Since a nationally representative sample of accidents was not feasible, areas were chosen to facilitate the development of models to predict injury severity from crash severity. In the development of these models it is necessary to obtain data that reflects the spectrum of accident experience of areas throughout the country. A model is only "representative" if it is applicable and predicts well for a wide range of areas with different environments. The areas chosen for NCSS were chosen to reflect some diversity in both urbanization and region of the country. Table 4.2 shows which areas represent different levels of urbanization and different regions of the country.

There is very little census information that is directly applicable to evaluating areas in terms of their accident populations. Measures like traffic density, number of registered vehicles, or drivers or miles of interstate highway are not readily available for all of the NCSS areas. The task is confounded by the fact that characteristics of the

---

[33] C. J. Kahane, R. A. Smith, and K. J. Tharpe (sic), "The National Crash Severity Study," International Technical Conference on Experimental Safety Vehicles. Sixth Report (Washington, D.C.: National Highway Traffic Safety Administration, 1978), pp. 493-516.

TABLE 4.2

Distribution of Sites by Region
and Degree of Urbanization

Area	North East	North Central	South	West
Central Cities[a]			Miami	Dynamic Science
Suburbs[b]	Calspan			
Other SMSA's[c]			SwRI(Urban)	
Non SMSA's or Small SMSA's[d]		HSRI Indiana	Kentucky SwRI:(Rural)	

[a]Central cities of Standard Metropolitan Statistical Areas (SMSA's) containing more than 1,000,000 persons

[b]Suburbs of SMSA's with more than 1,000,000 persons

[c]SMSA's with more than 250,000 persons but less than 1,000,000 persons

[d]SMSA's with fewer than 250,000 persons

people who live in the area of investigation may not represent the characteristics of the people who have accidents in that area. Thus the fact that areas exhibit similarities based on the characteristics of their residents may not imply that the areas do indeed have similar accident populations.

The County and City Data Book was used to obtain the following data on the areas chosen for NCSS. The data are presented in Tables 4.3 to 4.5. Table 4.3 gives general demographic information. Table 4.4 gives data on the persons living in the seven areas. Table 4.5 contains information that might have some correlation with the number of drivers in that area.

In viewing the tables it can be seen that for some characteristics the area percentages for all areas are quite comparable to the

TABLE 4.3

Site Characteristics - Demographic
Based on 1970 U. S. Population

Team	Land Area	Population Density	Death Rate	Percent Farm	Percent Urban	Percent Rural
Calspan	1016.7	640.00	7.4	1.0	79.0	19.8
HSRI	1464.0	215.65	6.8	5.1	68.5	26.5
U of Indiana	7165.0	69.77	10.2	10.5	39.8	49.6
U Miami	34.3	9769.00	13.9	0.0	100.0	0.0
U of Kentucky	1939.0	154.12	8.4	8.6	73.8	17.6
SwRI	13263	79.83	8.1	2.8	85.7	11.4
Dynamic Science L.A. County	19.5	40300.70*	9.3	0.0	100.0*	0.0
U.S.		57.46	9.5	4.1	73.5	22.4

*Based on 1977 population of the three police districts covered by Dynamic Sciences, Inc.

TABLE 4.4

Site Characteristics - Population

Team	Based on 1970 U.S. Population							Based on 1970 Civilian Labor Force		Based on 1970 Labor Force
	Percent Female	Percent Negro	Percent Spanish	Percent White	Percent Under 5 Years	Percent Over 16 Years	Percent Over 65 Years	Percent Female	Percent Unemployed	Percent White Collar
Calspan	51.1	0.7	0.29	98.8	8.5	63.2	7.8	35.0	3.8	64.2
HSRI	50.7	5.9	1.85	93.3	8.6	68.1	6.7	41.5	5.1	55.6
U of Indiana	51.1	0.5	0.17	99.0	8.4	66.0	10.6	36.2	4.6	38.7
U of Miami	53.3	22.8	45.3	76.7	6.3	74.6	14.5	46.1	4.3	40.9
U of Kentucky	51.6	10.5	0.28	85.5	8.4	67.8	8.7	40.0	3.3	49.1
SwRI	50.8	6.2	42.8	93.1	9.1	62.7	8.7	38.7	4.0	49.8
Dynamic Science L.A. County	24.3*	16.6*	41.1*	31.2*	8.2	68.3	9.5	39.4	6.4	55.8
U.S.	51.3	11.0	4.6	87.6	8.4	65.6	9.9	38.1	4.4	48.3

*All four percentages based on 1977 population of the three police districts covered by Dynamic Sciences, Inc.

TABLE 4.5

Site Characteristics - Automotive Related

Team	Based on 1970 U.S. Households		Based on 1970 Employed Civilian Labor Force		Based on Total 1970 Retail Sales		Based on Total 1970 Receipts	Percent Spent on*
	Percent with more than one Auto	Percent Low Income	Percent Use Public Transport	Percent work outside county of residence	Percent from Auto Dealers	Percent from Gas and Service Stations	Percent from Auto Repair Service	Highways
Calspan	92.1	4.1	4.8	5.6	18.7	7.1	10.4	7.8
HSRI	91.4	5.6	2.3	13.0	19.8	7.3	9.1	8.4
U of Indiana	86.1	10.0	0.8	20.4	19.2	7.8	3.5	11.2
U Miami	71.5	16.4	17.1	2.3	17.6	4.8	10.1	5.4
U of Kentucky	84.0	12.6	3.5	13.7	18.3	8.9	10.6	3.4
SwRI	85.6	17.2	4.8	6.8	20.7	8.7	11.1	6.2
Dynamic Science L.A. County	83.4	8.7	6.7	2.5	17.8	7.0	9.1	7.0
U.S.	82.5	10.7	8.9	17.8	17.9	7.9	11.6	7.6

*Local government finance, direct general expenditures.

percentage for the U.S. population, for example, the percentage female or the percentage unemployed. But, there are examples to the contrary. The percentage of Spanish population for the aggregate of the areas will be larger than the proportion in the total population. The proportion of the population that is White seems to be slightly overrepresented in the aggregate of the seven areas. The differences in these percentages may affect different phenomena differently. These differences may influence, for example, different crash types, in that a particular crash type may be over or under represented. There is no way to guarantee that the NCSS data is "nationally representative" but the data presented in Tables 4.3 to 4.5 suggest that on the average the areas when aggregated appear close to the national description for most variables considered.

4.2.2 Sources of Missing Data. There are three basic types of missing data that can occur in the NCSS data structure. An entire accident can be missed by the sampling system. This type of missing data is discussed in this subsection. If complete information about a vehicle or occupant is missing that will be referred to as unit non-response. Item non-response will refer to partial missing data. These latter two types of missing data are discussed together in Section 4.5.

There are three situations that cause accidents to be omitted from the sampling system. The first situation where accidents will be missing will only occur in the Phase 1 data relative to the Phase 2 sample design. In Phase 2 the definition of the population of accidents that were being sampled changed. The definition was expanded to include light trucks and vans. In the NCSS design there is a restriction on accidents that can be sampled. An accident is eligible to be in the sample only if the most severe injury occurred in a case vehicle. With this restriction it is possible that accidents which did not qualify in Phase 1 might have qualified for inclusion in Phase 2.

When combining the data from Phases 1 and 2 the population of accidents to described must be defined. The NCSS Statistics[34]

---

[34]Leda Ricci, ed., NCSS Statistics: Passenger Cars, Report No. UM-HSRI-80-36. Sponsored by the National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C., under

publication describes 27 months of accidents involving a towed passenger car. This task required effort to obtain the right subset of the Phase 2 data to be combined with the Phase 1 data. Simple aggregation of the 27 months of data will produce an underestimate of accidents involving towed passenger cars, light trucks and vans and an overestimate of accidents involving only passenger cars.

Accidents can also be lost at the time of collection, that is, some accidents may have been overlooked. It is also possible that accidents were collected on an incorrect day therefore losing the accidents that legitimately belonged in the sample. There is no way the data can be checked to see if all appropriate accidents were sampled. But due to random error some discrepancies may occur. In the process of calculating sampling errors it was found that for those teams sampling days and investigating all accidents on that day there were errors made. Accidents were included in the data that did not occur on legitimate sampling days. Table 4.6 summarizes the number of accidents found sampled incorrectly. Relatively, these are only a few cases but it does indicate that there is some variability due to collecting the appropriate accidents.

The most significant evidence of missing accidents was obtained in a comparison of NCSS, FARS, and various state files containing police-reported accidents. In this investigation, subsets from these files were created including only fatal occupants of passenger cars in NCSS areas. This analysis was done only on the Phase 1 data. The matching criteria used were subjective. Comparisons between cases were based on the date of the crash, county, age and sex of the fatality, time of the accident, and the total number killed in the accident. If information on one or two of the variables listed above was different in the NCSS, FARS, or state file, and the discrepancy was not serious the cases were matched. If the discrepancy was serious, such as the dates differing by a couple of weeks, the vehicle makes of the passenger cars involved in the crash were compared to make sure it was the same case. Cases with serious reporting discrepancies were matched. For example, in Michigan

---

Contract No. DOT-HS-8-01944 (Ann Arbor: Highway Safety Research Institute, The University of Michigan, June 1980).

TABLE 4.6

Ineligible Accidents Sampled

Team	Number of Accidents	
	January 1977 - March 1978[a]	April 1978 - March 1979[b]
Calspan	5	11
Indiana	6	4
Kentucky	6	0
Miami	3	2
Los Angeles	11	9
Total	31	26

[a]Calspan has verified that these accidents were sampled on the wrong day.

[b]These accidents have not been verified by Calspan. They may include some cases that have an error in coding the date of the accident.

a fatality had two reported dates differing by 16 days. In this case, all other variables were consistent between NCSS and FARS so the case was considered matched.

A mis-match was defined to be any case listed in one file that was not listed in the other. In Michigan and Texas a mis-match was a case listed in only one of the three comparison files. It should be noted that the populations described by NCSS and FARS are not totally compatible. Inclusion in NCSS was determined by the worst injury in a towed vehicle. FARS includes all fatalities independent of whether the vehicle was towed. Accidents where a fatality occurred after the crash but where the accident was not the cause of death are also included in FARS, and these cannot be excluded from analysis. These

incompatibilities suggest that there may be fatal cases in FARS that by definition should not be in NCSS.

The results of this analysis are presented in Table 4.7. These figures indicate that there were fatalities missed in both the NCSS data collection and the FARS data collection. Of the total number of fatalities found 93.4% were represented in FARS and 81.9% were found in NCSS. This gives an approximation to how extensive the problem of under-coverage of fatalities is in NCSS. Again, this number is only approximate since the total number of fatalities found may include fatalities not eligible for inclusion under the NCSS design.

TABLE 4.7

Undercoverage of Fatal Accidents
Phase 1

Team	Total Fatalities Found	Fatalities Found in NCSS	Fatalities Found in FARS	Matched Fatalities
Calspan	61	53	61	52
Kentucky	55	49	52	46
Miami	31	21	29	19
HSRI	62	51	58	46
Indiana	110	101	99	90
SwRI	179	133	172	130
LA*		19	19	
Total	498	408	470	383

*In Los Angeles all fatalities in NCSS were found in FARS. The fatalities for the specific part of Los Angeles could not be determined in FARS.

4.2.3 Summary. In general the areas chosen for inclusion in NCSS appear quite similar to the U.S. population distribution. There are some characteristics that appear to be overrepresented in the NCSS areas that may indirectly result in the overrepresentation of an accident related characteristic. The distributions presented do not address directly whether the accident population in the NCSS areas is comparable with the U.S. accident population. Information for such an analysis is not available. These seven areas appear to give a firm base to the modelling efforts described in Section 3.

In analyzing the Phase 1 and Phase 2 data together it should be noted that the data was obtained sampling two different accident populations. Phase 1 data represents accidents involving a towed passenger car. The Phase 2 data expands the accident population to all accidents involving a towed passenger car, light truck or van. The result is that the number of accidents involving a towed vehicle sampled in Phase 1 is less than the number of accidents involving a towed vehicle that would have been required under the Phase 2 design. The key observation is that the Phase 1 and Phase 2 data are samples from different accident populations.

The estimation of statistics for the aggregate depends on accurate collection of all accidents sampled under the sample design. It is expected that a slight variability is present due to the sampling process. A more serious source of undercoverage was found in the the number of fatalities investigated. A census of all fatalities was required by the sample design and by matching these fatalities with FARS it is possible that as much as 20% of the fatalities were not investigated.

For most estimated accident statistics for the NCSS areas, adjustment for missing fatal data will not change the estimate appreciably. One important exception is the distribution of OAIS. The frequencies at the more severe end of the OAIS scale would be underestimated due to the missing fatal data. If the assumption that the distribution of OAIS for the unobserved fatals is the same as for the fatals observed is reasonable, an adjustment to the number of fatalities in each OAIS category can be made. This adjustment inflates

the totals by the response rate, that is, the rate of observing non-missing fatal occupants. These adjusted totals will provide a modified distribution for OAIS. This estimated distribution will be sensitive to the estimate of the number of fatals missing.

Missing data due to missing accidents from the sample is the hardest problem to address statistically since there is usually no information about these accidents or an exact count of the number missing. The solution to the problem is a dedicated effort to make sure every accident that is designated to be in the sample is represented in the data collected.

4.3 NCSS Statistics

During this project four publications were produced to present summary statistics from NCSS. These four reports are:

1. NCSS Statistics: January 1977-March 1978, October 1979[35]
2. NCSS Statistics: Light Trucks and Vans (Preliminary), December 1979[36]
3. NCSS Statistics: Passenger Cars, June 1980[37]
4. NCSS Statistics: Light Trucks and Vans, June 1980[38]

These publications present in a concise form accident statistics for accident, vehicle, and occupant variables collected in NCSS. Each publication is organized into five major sections. The first presents a general overview of the accidents described by the particular publication. In the second section, statistics on the accidents are presented. The third and fourth sections provide information on the vehicles and occupants respectively. The last section presents various collision severity (Delta V) distributions. The first publication describes Phase 1 data (January 1977 to March 1978). The remaining publications describe the Phase 2 data (April 1978 to March 1979). The

---

[35] Leda Ricci, ed., NCSS Statistics: January 1977-March 1978, Report No. UM-HSRI-79-80. Sponsored by the National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C., under Contract No. DOT-HS-8-01944 (Ann Arbor: Highway Safety Research Institute, The University of Michigan, October 1979).

[36] Leda Ricci, ed., NCSS Statistics: Light Trucks and Vans (Preliminary), Report No. UM-HSRI-79-95. Sponsored by the National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C., under Contract No. DOT-HS-8-01944 (Ann Arbor: Highway Safety Research Institute, The University of Michigan, December 1979).

[37] Leda Ricci, ed., NCSS Statistics: Passenger Cars, Report No. UM-HSRI-80-36. Sponsored by the National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C., under Contract No. DOT-HS-8-01944 (Ann Arbor: Highway Safety Research Institute, The University of Michigan, June 1980).

[38] Leda Ricci, ed., NCSS Statistics: Light Trucks and Vans, Report No. UM-HSRI-80-37. Sponsored by the National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C., under Contract No. DOT-HS-8-01944 (Ann Arbor: Highway Safety Research Institute, The University of Michigan, June 1980).

second publication listed is a preliminary version of the fourth publication listed.

In each publication, for the most part, tables are presented in two complementary forms. One page provides a frequency distribution of the factor under consideration; the opposing page shows the corresponding injury rates. In each case the columns of these tables show the number of occupants in each of several injury categories. These categories were defined by the NCSA generated variables NEWOAIS2 and NEWOAIS3.[39] These variables are defined using information from various injury variables, including the Abbreviated Injury Scale[40] (AIS) and have less missing data than the Overall AIS.

In the publications that describe passenger cars there was extensive graphical displays of the data presented in the tables. Graphical methods include histogram, bar graphs, pie graphs, line graphs, and three dimensional plots. The method of graphical display was chosen to complement the data presented in the table.

All of the statistics were developed using the estimation methods described in Section 4.1.2. Sampling errors were not included in any of these publications. In Section 4.4 the problem of presentation of sampling errors for a large scale data collection effort is discussed. The sampling errors associated with selected accident, vehicle and occupant level accident statistics are also presented there.

For passenger cars Phase 1 data represents a smaller accident population that the Phase 2 data. It should be noted that in order to produce a publication that describes the combination of the Phase 1 and Phase 2 data the Phase 2 data need to be subset. This subsetting procedure will now be described.

The requirement here was to eliminate those accidents that would have been excluded in the last twelve months of NCSS under the criteria of the first fifteen months of the study. During the first fifteen

---

[39] These derived variables are documented in Appendix A.

[40] The Abbreviated Injury Scale (Morton Grove, Ill.: American Association for Automotive Medicine, 1976).

months only passenger cars were eligible for selection as case vehicles. Each accident had to include at least one case vehicle and no accident could be selected unless the most serious injury in the accident occurred in a case vehicle. In the last twelve months of the study, however, light trucks and vans could be case vehicles.

The hierarchy of seriousness of injury was based on the "NCSS Classification":

Most Serious: Fatal injuries and overnight hospitalization.

Next Most Serious: Transported to hospital and released.

Least Serious: Any other treatment.

If all the necessary data had been complete one simple algorithm would have sufficed for the production of the passenger-car subset:

> Identify all accidents in which the most serious injury (NCSS Classification) occurred in a passenger car.

This would involve identifying for each case vehicle the most serious injury to an occupant and then identifying within each accident the vehicle with the most serious injury. If the vehicle with the most serious injury were a passenger car, or injuries of equal severity occurred in a passenger car and a light truck or van, then the accident would be included. If the most serious injury occurred in a light truck or van the accident would be excluded.

This procedure was followed and a total of 4669 accidents (out of an original total of 5305) were identified for inclusion in the subset.

However there were 34 vehicles missing the body-type variable. It was decided to count these vehicles as passenger cars. This resulted in an additional 27 accidents identified for inclusion.

Finally there was the problem of missing data on injury severity. If the data was missing for any occupant of a vehicle it was impossible to calculate the most severe injury in a vehicle and therefore in the accident concerned. But if an accident involved no case vehicle light trucks or vans then that accident should be included in the subset anyway. This procedure identified a further 7 appropriate accidents, resulting in a grand total of 4703 accidents in the subset.

The procedure could also have been reversed so that first accidents involving no case vehicle light trucks or vans would have been identified, and second, accidents where the most serious injury was in a passenger car. Under this procedure 4400 accidents would have fallen in the first category, 303 in the second, thus once again resulting in 4703 selected accidents.

The procedure for producing the appropriate vehicle, occupant, case vehicle and case vehicle occupant files was essentially the same as that used in subsetting the truck accidents, involving a match on the accident ID's between the newly created subset and the full vehicle or occupant files.

The final step was to combine the files (which now contained only a limited number of variables) with the corresponding file from the first fifteen months of the study, producing five datasets, each covering the full 27 months.

The publication of the statistics for light trucks and vans depended on subsetting the Phase 2 data to obtain the appropriate set of accidents. This was accomplished in the following manner.

The goal in producing the truck subset was to include only those accidents from the last twelve months of NCSS that involved a towed light truck or van. An applicable vehicle, towed from the scene of an accident, is called a "case vehicle" in the study. During the first fifteen months of the data collection only passenger cars were applicable vehicles, but for the last twelve months light trucks and vans (passenger vans, cargo vans, pickups etc.) were also counted as applicable vehicles.

The first requirement was to identify the appropriate accidents. This was done by matching the vehicle file with the accident file. A variable was created in the accident file that identified all accidents involving towed light trucks or vans. A new accident file was then created including only the appropriate accidents. There were 905 cases in this file.

A match was then performed between the newly created accident file and the full vehicle file to identify all the vehicles in the

appropriate accidents. 1563 vehicles were identified and a new vehicle subset was produced, containing both accident and vehicle variables.

A similar match was made with the full occupant file and a subset of occupants was produced. This new file contained 2514 occupants.

Finally a vehicle file that included only towed light trucks and vans (a case vehicle file) and a file for occupants of those vehicles (a case vehicle occupant file) was produced. These last two datasets contained 951 and 1608 cases respectively.

In summary, there are some key points to be made concerning these publications. NCSS fails to satisfy the requirements of a national probability sample. The areas chosen for data collection were not selected randomly. They were chosen to be geographically diverse. The areas were also selected so that the distribution across urbanization types for these areas was approximately the same as the U.S. distribution. A method for generating "national projections" is described in Section 4.1.5 and is used for various NCSS statistics in Section 4.5. The statistics in these publications describe police-reported accidents involving towed appropriate vehicles for the aggregate of the seven areas.

For many variables there are substantial portions of missing data. Missing data counts and percentages are shown for all row variables. Adjustment for missing data is discussed in Sections 4.1.4 and 4.5. It should be noted that missing data counts and percentages have not been excluded in the calculation of column percentages and, consequently, the percentages shown may be slightly underestimated.

The total number of light trucks and vans described by the light truck and van publication is only about 5% of that for the passenger cars represented in the final two publications. The data collection for the passenger car publication covered 27 months while data collection for light trucks and vans covered only the last 12 months of that period. The light trucks and vans are distributed among the seven areas far less uniformly than the passenger car accidents. This factor suggests that the accident experience of light trucks and vans in the

aggregate of the seven areas is not described nearly as well as the accident experience of passenger cars.

In general, the tow-away accident population for light trucks and vans is not directly comparable with that of the passenger cars. In particular, the light truck and van accidents tend to occur at greater crash severities. The proportion of light truck and van accidents occurring in rural areas is about 76% greater than that of passenger cars. Even the "tow-away" threshold is likely to be different for trucks as compared to passenger cars.

4.4 Precision of Estimates

In this subsection sampling errors are discussed for accident, vehicle, and occupant level NCSS proportions. These sampling errors are calculated according to the NCSS sample design which is basically a stratified cluster sample within each area chosen. No attempt was made to calculate all possible sampling errors but 19 proportions were chosen to be representative and sampling errors were calculated for these proportions.

These sampling errors are summarized using two techniques. In the first, estimated proportions are graphed as a function of the estimated sampling errors. In the second, variables with the same design effect are grouped to give a description of variables that have the same adjustment to the simple variance to account for the NCSS design. Design effects for NCSS proportions are described in this subsection. These design effects are helpful in summarizing the magnitude of the sampling errors for specific types of statistics when data is collected from the same design over a long period of time.

4.4.1 Variance Estimation. Proportions and variances of proportions were calculated for nineteen statistics. These nineteen statistics were calculated separately for each design group.[41] The nineteen proportions were broken down as follows:

3 Accident Proportions:
- Proportion of rural accidents
- Proportion of accidents during rush hour
- Proportion of accidents on dry roads

8 Vehicle Proportions:
- Proportion of vehicles with front CDC
- Proportion of vehicles with right CDC
- Proportion of vehicles with back CDC
- Proportion of vehicles which underwent intrusion
- Proportion of vehicles which did not undergo intrusion
- Proportion of vehicles with low Delta V
- Proportion of vehicles with high Delta V
- Proportion of vehicles which hit another car

---

[41] Design groups are defined in Section 4.1. Some areas used different systematic samples for specific counties within the area. The ten design groups are described in Table 4.1.

8 Occupant Proportions:

- Proportion of occupants aged 16 and under
- Proportion of occupants aged 17 through 30
- Proportion of occupants aged 31 through 45
- Proportion of occupants aged 46 and over
- Proportion of occupants not wearing a seat belt
- Proportion of occupants wearing a seat belt
- Proportion of occupants with OAIS 0 through 2
- Proportion of occupants with OAIS 3 through 6

The last two proportions were taken from the variable NEWOAIS3. This variable is calculated with an NCSA designed algorithm and is an attempt to overcome some of the missing data problems with AIS. Documentation on this algorithm is presented in Appendix A.

The proportions were selected in the expectation that some of them might show quite large effects from the cluster design, but others would be less susceptible.

The estimated probabilities and variances that were calculated are presented in Tables 1 to 3 of Appendix C. For those design groups using a cluster design (i.e., all proportions except those for HSRI and SwRI at the accident level) the estimated probabilities were calculated using Equation 4-4. The estimated variance of these proportions is given by Equation 4-13. The probabilities and variances of the accident proportions for the two design groups that sampled accidents were calculated using Equations 4-4 and 4-10.

All of the sample designs used in the NCSS involve a cluster sample, either accident or day. When sampling clusters, the clusters become the basic unit of observation and cluster totals are used in the calculation of the variance of statistics estimated. The variance estimation formulae are presented and discussed in Section 4.1.3. To use these formulae either a program must be written to calculate the appropriate variance or a new data structure can be created. Once this data structure is created the variance estimation procedure becomes straight forward. An algorithm for creating the new data structure representing cluster totals is given in Appendix B.

An overview of the estimated variances of the proportions is presented in Table 4.8. In this table, as in most of those that follow, the accident-level proportions for HSRI and SwRI are not included as

they were not sampled using a cluster design and are therefore irrelevant to a discussion of how the estimates are affected by cluster designs. Where these proportions are included, this will be noted. Table 4.8 gives the mean estimated variance for the proportions calculated by variable type. For example, the three accident proportions for each of the 10 design groups have a mean estimated variance of .001911. The largest estimated variance for these 30 proportions was .00707. The standard deviation of these 30 estimated variances is .00171.

TABLE 4.8

Overview of Estimated Variance of Proportions

Proportion Type	Sample Size	Minimum	Maximum	Mean	Standard Deviation
Accident Proportions	30	.0000159	.00707	.001911	.001710
Vehicle Proportions	80	.0000315	.00275	.000621	.000615
Occupant Proportions	80	.0000017	.00202	.000346	.000393

Table 4.8 shows that, as might be expected, the estimated variances are highest overall for the accident proportions, followed in turn by the vehicle and occupant proportions. The accident proportions are more susceptible to the effects of clustering since only a few accidents were available for selection by each design group on each sampling day. Frequently no accidents at all, or merely one or two accidents, were selected by a particular design group in the 25% or 10% sampling strata. The high estimated variances resulting indicate the instability of the probabilities estimated from a design where only a few cases per cluster are available. This was much less of a problem at the vehicle and occupant levels where the much larger number of cases made the problem of producing stable estimates less acute.

A further indication of the increasing stability of the estimates with the increase in the number of cases is given by the decline in the standard error of the estimated variance, from accident to vehicle to occupant proportions, as shown in the right-most column of Table 4.8. Not only did the estimated variance of the estimated proportion decline, but these variances became consistently smaller. This means that at the occupant level the variances were generally small, while at the accident level they were large overall, but not consistently so.

In the discussion thus far only one source of error has been dealt with. This source of error resulted from the variance of a proportion within design groups. There is however a further source of error which occurs when averaging proportions to produce estimates of proportions for the aggregate, averaging over all design groups. Such a proportion is produced when an estimate is made of the overall proportion of rural accidents in all the study areas. This further source of error is the variance between design groups. A rough approximation of the overall between-design group error is given in the first two columns of Table 4.9. These between-design group errors were calculated as simple variances of a given proportion across design groups. In calculating these figures and all the others in the table, the accident-level proportions for HSRI and SwRI were included.

The two right-hand columns present the simple mean of the within-design group estimated variances calculated across the two or eight appropriate design groups. Looking at these two right-hand columns first, the mean variance for the two design groups that sampled by accident is invariably lower than the mean estimated variance for the eight design groups that sampled by day. The proportions calculated using a sample of accidents are considerably more stable estimates than the proportions calculated using a sample of days.

The between errors are, not surprisingly, greater at the accident level, where environmental factors such as local climate or degree of urbanization are more likely to influence the proportions. There is no discernible pattern to the size of the between error by sample design, nor should one be expected. Error is not invariably larger between the design groups that sampled by day than it is between the groups that

TABLE 4.9

Between and Within Design Group Variance
by Sample Design

Proportion Type	Proportion	Variance of Proportions Between Design Groups		Mean Variance of Proportion Within Design Groups	
		2 Groups Sampling by Accident	8 Groups Sampling by Day	2 Groups Sampling by Accident	8 Groups Sampling by Day
Accident	Rural	.039000	.097000	.000360	.001100
	Rush Hour	.000000	.026000	.000400	.001400
	Dry Road	.041000	.021000	.000400	.003100
Vehicle	Front CDC	.000730	.006400	.000280	.000990
	Right CDC	.000006	.000550	.000140	.000310
	Back CDC	.000037	.000130	.000043	.000260
	Intruded	.000440	.006600	.000190	.000600
	Not Intruded	.000006	.007200	.000290	.000970
	Low Delta V	.000950	.005800	.000380	.000740
	High Delta V	.002900	.008200	.000240	.000710
	Hit a Car	.004600	.011000	.000430	.001100
Occupant	Aged 16 or Under	.001300	.002000	.000210	.000440
	Aged 17 to 30	.001900	.001400	.000370	.000830
	Aged 31 to 45	.000093	.001200	.000110	.000240
	Aged 46 and Over	.000032	.001600	.000170	.000480
	Unbelted	.002500	.013000	.000280	.000720
	With OAIS 0-2	.001700	.015000	.000180	.000240
	With OAIS 3-6	.000001	.000360	.000004	.000025

sampled by accident. Between error represents real differences between design groups and is not a function of the design.

For some proportions at least, between error is larger than within error. Examples of this are rural accidents, dry road accidents, front CDC and Delta V proportions and OAIS between 0 and 2. Thus if only within errors are taken into account (as they are below in calculating the design effects) the true variance of a proportion will be underestimated. This underestimation will sometimes be considerable if the omitted variance is larger than the included variance.

4.4.2 Graphical Presentation of Estimated Variances. It was desirable to obtain an idea of how the variances of the estimated proportions were related to the estimated proportion. Such relationships could be used in predicting the variances for other proportions beyond those calculated here. For this purpose, plots, one for each design group, were made to show the distribution of the estimated variance of the estimated probability against the estimated probability. Such a distribution should be parabolic. Small and large proportions normally have small variances, while as the proportion approaches .5 the variance should increase to its maximum. Indeed the general distribution for each design group did follow the expected pattern. Two of the distributions are presented in Figures 4.1 and 4.2. In both cases a line approximating a parabola could be drawn in such a way as to minimize deviations from it, leaving one two proportions as outliers.

Perhaps of more general use are the distributions shown in Figures 4.3 through 4.5. Here the mean of the estimated proportions for each proportion across design groups is plotted against the mean of the estimated variance of the proportion across design groups. Each distribution follows the expected parabolic curve, though this should become more apparent if further points on the graphs were calculated. These distributions could be used to predict, for any proportion, the approximate size of the average within-design group variance. Thus an estimate of this source of error could be made for any proportion

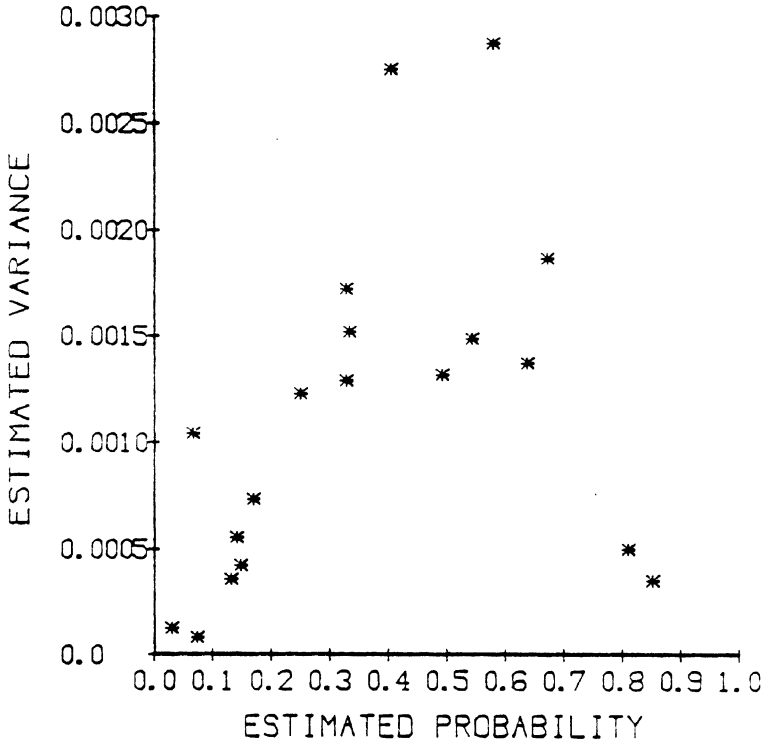

FIGURE 4.1 Estimated Probabilities and Estimated Variances for Indiana B

calculated for NCSS Statistics: January 1977 - March 1978.[42] The particular distribution referred to for the estimate would depend on whether the proportion was coded at the accident, vehicle or occupant level.

4.4.3 Design Effects. Design effects were calculated for all 190 design group proportions for which proportions and variances had been estimated. Design effects are defined in Section 4.1.3. No finite population correction factors were included and the denominator used was a simple random sample variance of unweighted cases that included the cases sampled with certainty.

The design effect shows the increase (or decrease) in variance compared to a simple random sample variance and in the case of this

---

[42]Leda Ricci, ed., NCSS Statistics: January 1977-March 1978, Report No. UM-HSRI-79-80. Sponsored by the National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C., under Contract No. DOT-HS-8-01944 (Ann Arbor: Highway Safety Research Institute, The University of Michigan, October 1979).

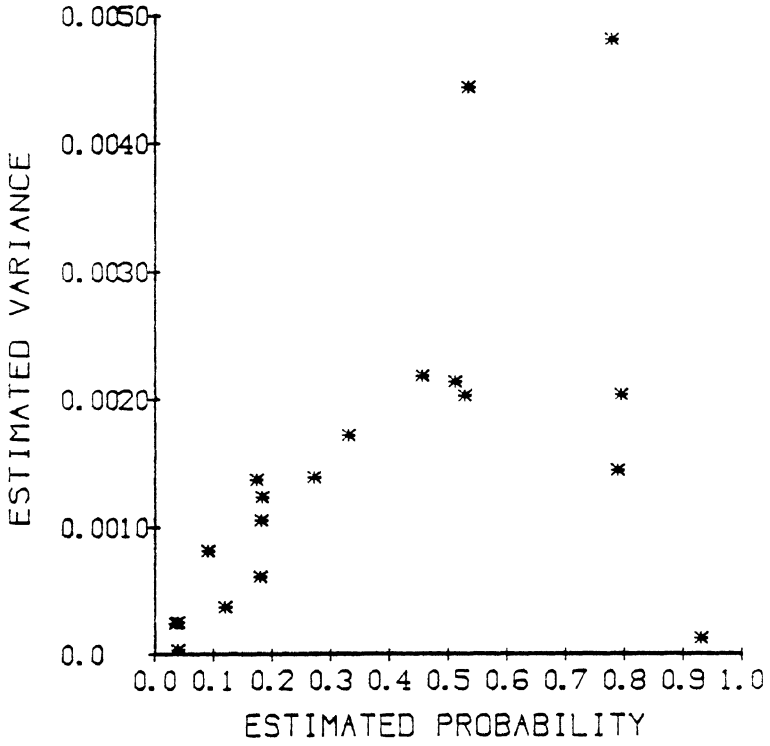

FIGURE 4.2 Estimated Probabilities and Estimated Variances for Kentucky B

study this increase is the effect of both stratification and clustering. The 190 design effects are presented in Tables 4 to 6 of Appendix C but they are summarized in Table 4.10. This table does not include the accident-level proportions for HSRI and SwRI. Table 4.10 gives the mean estimated design effect for proportions estimated by variable type. For example, the eight vehicle proportions for each of the 10 design groups have an mean estimated design effect of 1.836. The smallest estimated design effect for these 80 proportions is .584. The standard error of these 80 estimated design effects is 1.275.

Clearly the estimated design effects for the accident proportions were considerably larger than those for the vehicle and occupant proportions. The extreme case, with a design effect of 21 was the proportion of dry road accidents for the Miami data collection area. It should be borne in mind that the design effect represents a ratio of variances: the ratio of standard deviations is the square root of the design effect. The proportion of dry road accidents in Miami has a variance that is 21 times that it would have been from a simple random

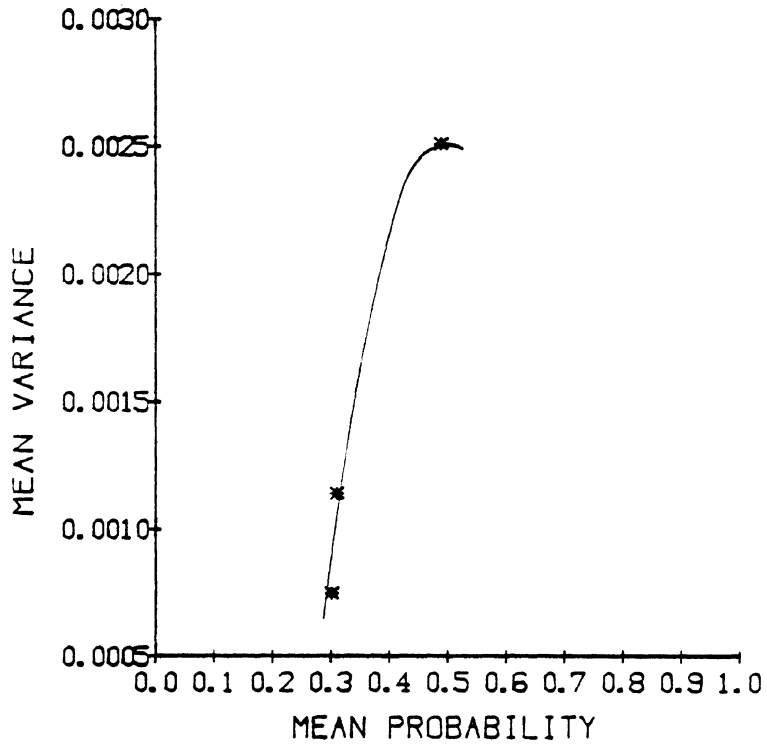

FIGURE 4.3 Mean Probabilities and Mean Variances for the Accident Proportions

sample of equal size. But the standard deviation and hence the confidence interval is 4.6 times what it would have been from a simple random sample, though this is still a very large factor.

Table 4.10 would appear to indicate once again that accident variables are far more susceptible to the influence of environmental factors than vehicle and occupant variables. The mean accident-level design effect is about twice as large as the mean vehicle- or occupant-level design effect.

Table 4.11 indicates which proportion, of those computed, is most responsible for the high mean design effect at the accident level. It is the proportion of accidents on dry roads, a statistic for which one would expect there to be a large clustering effect. On a particular day the weather is likely to be dry or wet for the whole day, so that all the accidents will occur on dry roads or wet roads as the case may be.

Table 4.11 also shows how the sample design influences the design effects. At the accident level the two groups that sampled by accident had design effects that were consistently smaller than the effects for

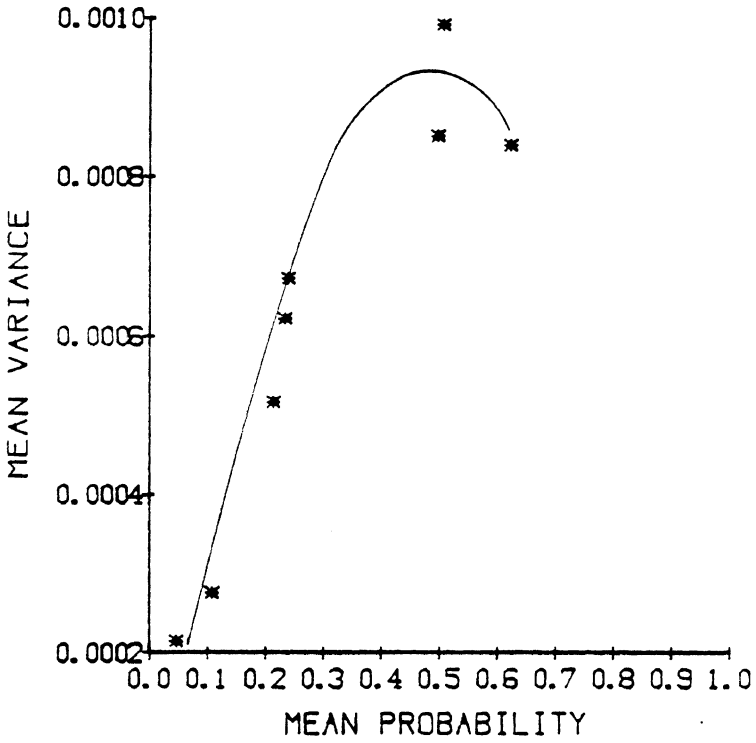

FIGURE 4.4 Mean Probabilities and Mean Variances for the Vehicle Proportions

the groups that sampled by day. This is particularly so for the proportion of accidents on dry roads. Overall, the mean design effect for the accident proportions was 1.623 for the groups that sampled by accident and 3.846 for the groups that sampled by day.

At the vehicle and occupant levels no such consistent pattern can be discerned. Back CDC stands out for having a much greater design effect when sampled by day. Other proportions would appear to show a clustering effect by accident (in reality more likely by vehicle) rather than by day.

Interestingly the smallest estimated design effects of all are those for the proportion of occupants with serious injuries (OAIS 3-6). The variable used to define the sampling strata is highly correlated with injury level. It is the NCSS treatment category of the occupant in the accident with the most "serious" treatment. So the very small design effects here show the benefits of sampling using a design that stratifies on a variable of interest, or one that is highly correlated with a variable that is being studied. It is striking that the design

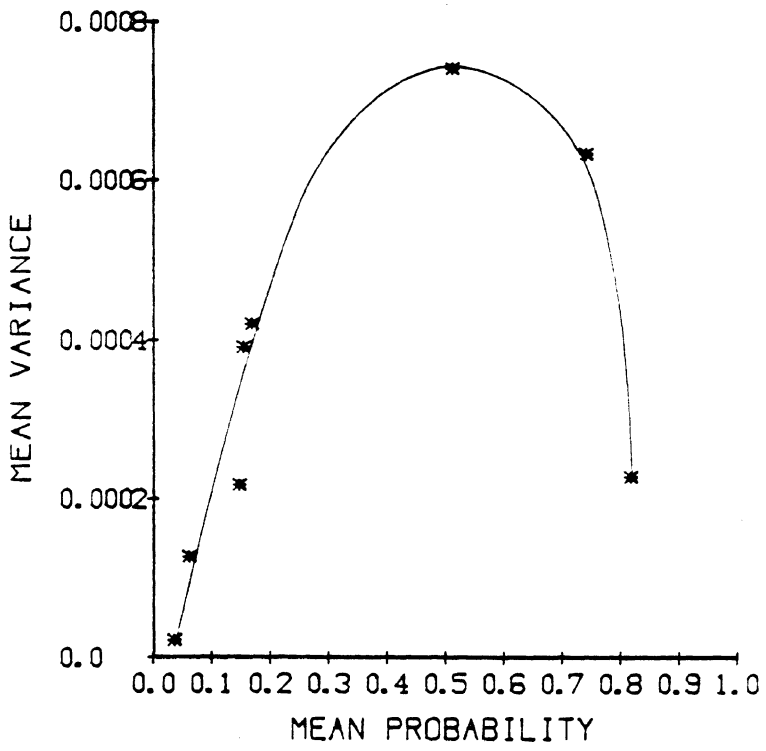

FIGURE 4.5 Mean Probabilities and Mean Variances for the Occupant Proportions

TABLE 4.10

Overview of Design Effects

Proportion Type	Sample Size	Minimum	Maximum	Mean	Standard Deviation
Accident Proportions	30	.970	21.142	3.846	4.187
Vehicle Proportions	80	.584	11.296	1.836	1.275
Occupant Proportions	80	.049	7.154	2.238	1.287

effects for the proportion of occupants with minor injuries (OAIS 0-2) have not been similarly affected. A possible explanation is that in accidents where one occupant receives severe injuries the other

TABLE 4.11

Design Effects by Proportion and Sample Design

Proportion Type	Proportion	Mean Design Effect	
		2 Groups Sampling by Accident	8 Groups Sampling by Day
Accident	Rural	1.293	1.994
	Rush Hour	1.816	2.168
	Dry Road	1.758	7.143
Vehicle	Front CDC	1.401	1.681
	Right CDC	1.563	1.359
	Back CDC	1.638	3.059
	Intruded	0.958	1.040
	Not Intruded	1.414	1.738
	Low Delta V	2.996	2.571
	High Delta V	1.435	1.434
	Hit a Car	2.045	2.115
Occupant	Aged 16 and Under	2.895	2.698
	Aged 17 to 30	2.921	2.641
	Aged 31 to 45	1.966	1.738
	Aged 46 and Over	2.494	2.614
	Unbelted	3.116	3.839
	Belted	3.033	3.036
	With OAIS 0-2	1.470	1.183
	With OAIS 3-6	0.102	0.137

occupants are just as likely to receive minor injuries as they are in accidents where no occupant is severely injured.

The overall mean of design effect at the vehicle and occupant levels are roughly the same for the two sample designs. The vehicle-level mean design effect is 1.681 for the sample of accidents and 1.875 for the sample of days. The mean design effect at the occupant-level are 2.250 and 2.236 respectively.

A visual representation of all the design effects calculated is given in Figures 4.6 through 4.9. From Figure 4.6 it can be seen how each individual design group statistic contributed to the high mean design effect shown for the proportion of accidents on dry roads in Table 4.11. The proportion for Miami has a design effect of 21.142.

Only the design groups that sampled by accident (HSRI and SwRI) have design effects lower than 2 for this proportion. Among the design group vehicle proportions, the proportion of vehicles with Back CDC for Indiana B stands out with a design effect of 11.296. Of the design group occupant proportions the most prominent is the proportion of occupants not wearing seat belts for Miami with a design effect of 7.154.

The clustering effect for Back CDC would appear to be explained in part by the correlation of Back CDC with accidents on non-dry roads (that is, on roads which are wet, snow-covered, or icy). In Stratum 3 (the 10% stratum) there was a correlation of .42 between the proportion of vehicles in the Calspan area with Back CDC and proportion of accidents on non-dry roads. The same correlation for the Kentucky B area was .47. The explanation would appear to be that there are more rear-end collisions on wet roads.

Finally, a graphical representation of the confidence intervals that resulted from the actual sample design as compared with those that would have resulted from a simple random sample of equal size is given in Figure 4.10. For the accident-level proportion shown (which had a design effect of 7.913) the 95% confidence interval stretches all the way from a proportion of .28 to one of .55. The 95% confidence interval from a simple random sample of equal size would have ranged from .38 to .47. For the vehicle and occupant proportions shown the difference in the confidence intervals between the two types of samples is negligible. In the case of the vehicle proportion the confidence interval from the stratified sample of days is actually reduced compared to that from the simple random sample.

It should be noted that the design effects presented in this section and the confidence intervals displayed in Figure 4.10 are for individual design group proportions. They are not for proportions aggregated across design groups to make totals for the whole study area. Such totals are affected by the further source of variance discussed in Section 4.4.1, the between design group variance which for some proportions was larger than the mean within design group variance. The

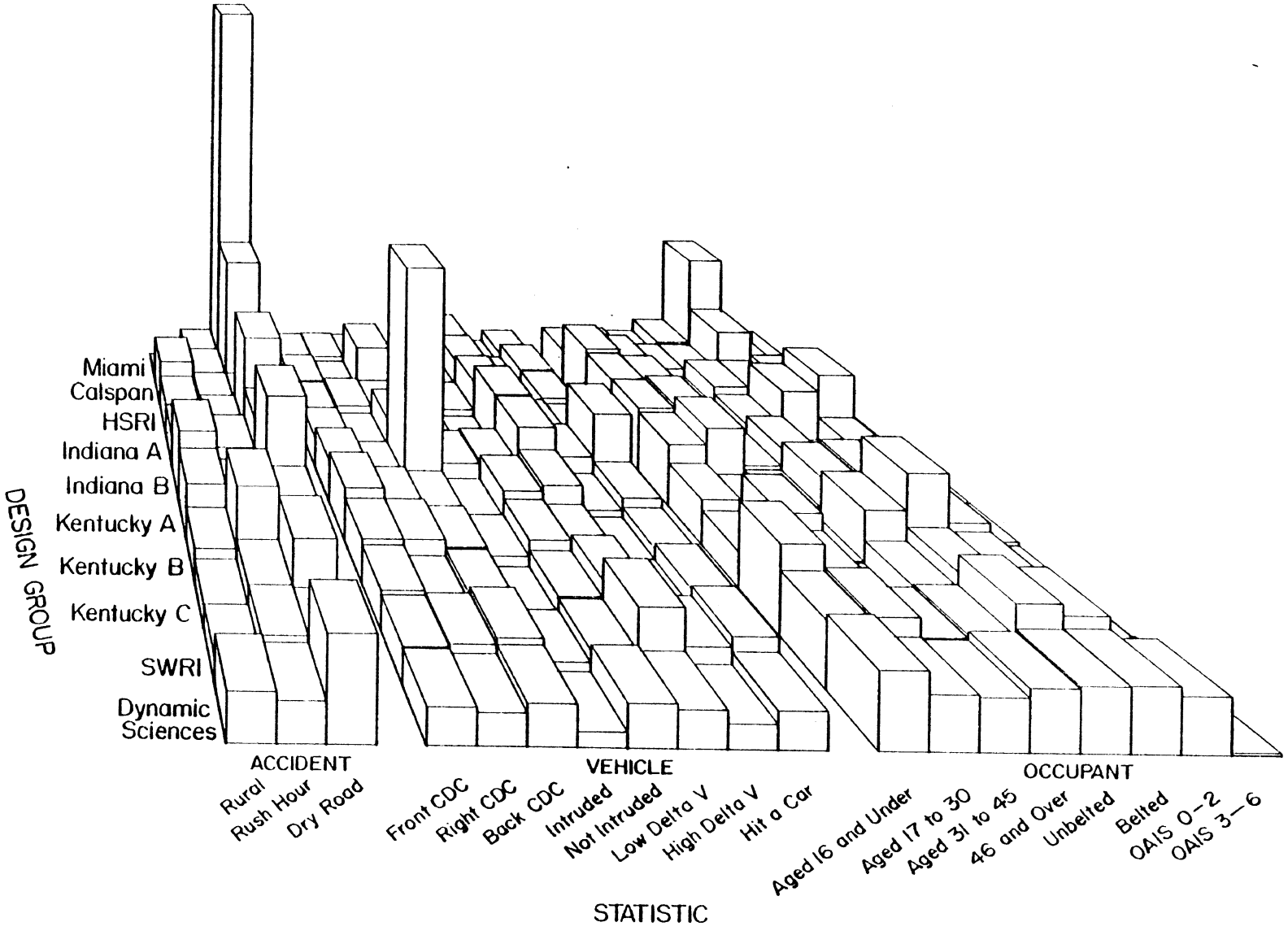

FIGURE 4.6 DESIGN EFFECTS BY DESIGN GROUP AND STATISTIC

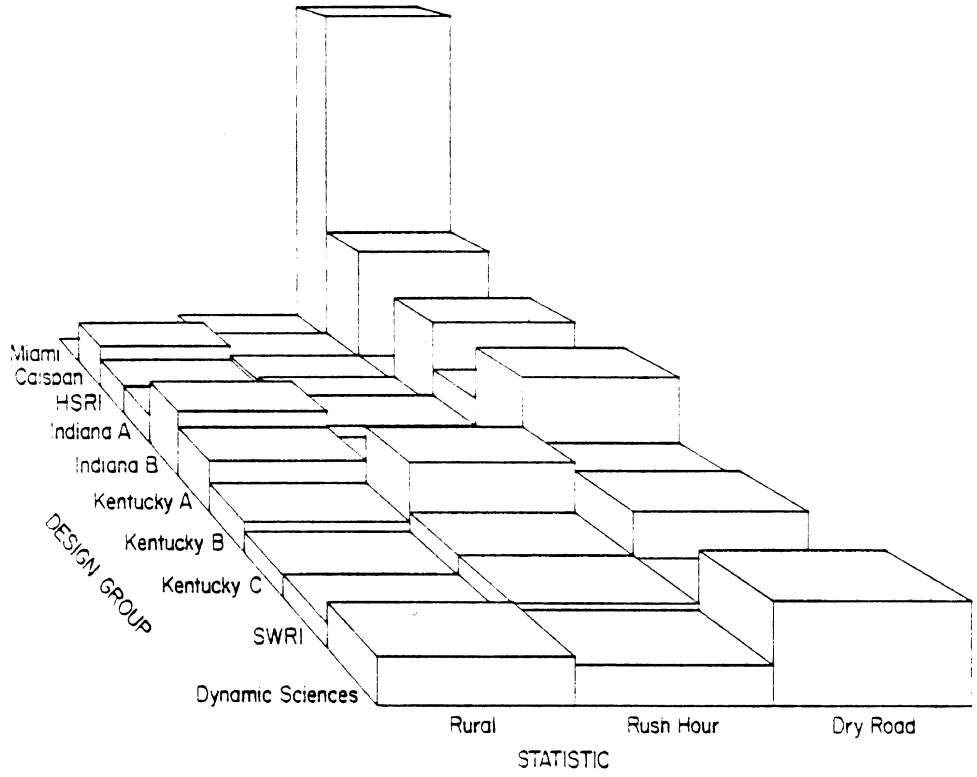

FIGURE 4.7 Accident Design Effects by Design Group and Proportion

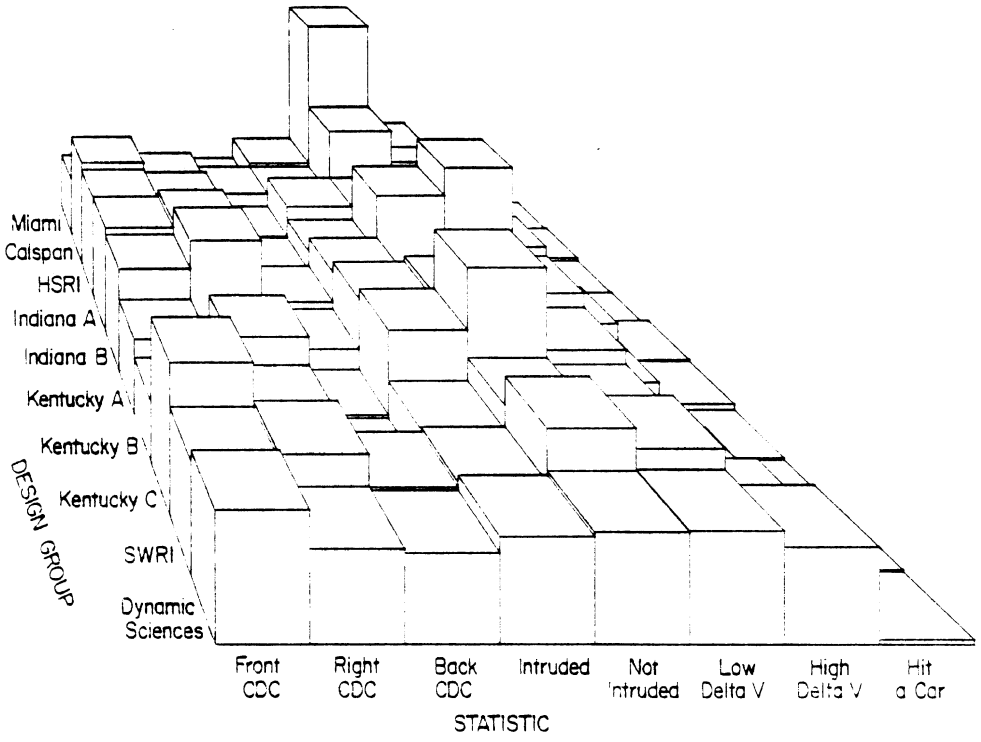

FIGURE 4.8 Vehicle Design Effects by Design Group and Proportion

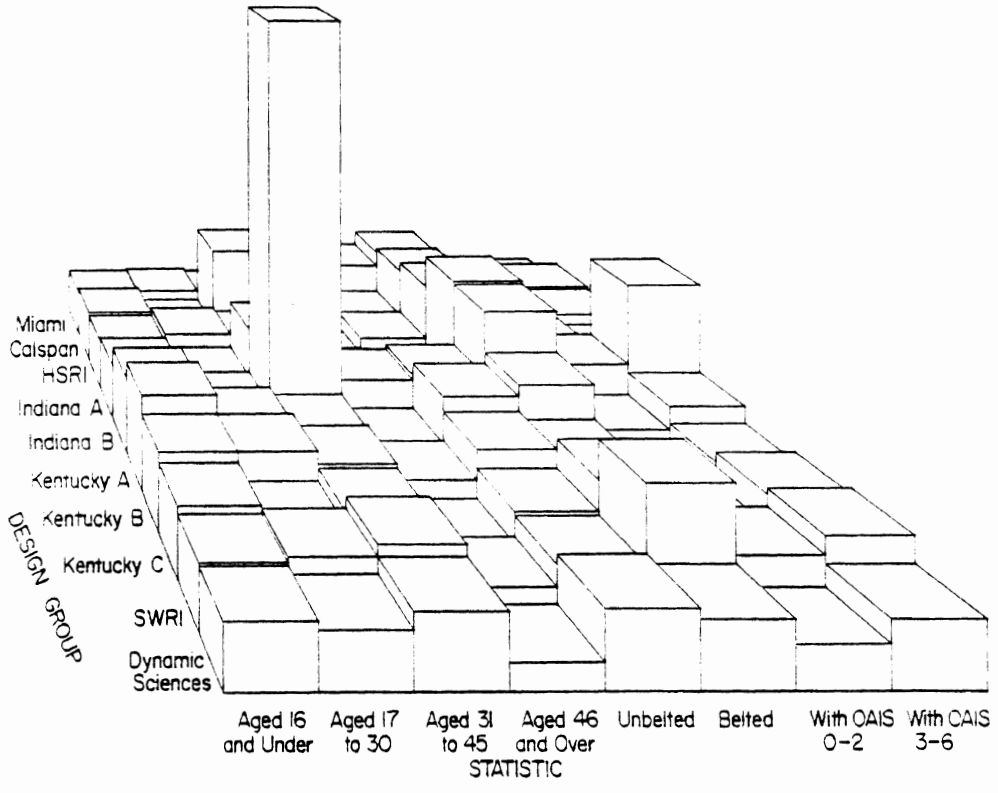

FIGURE 4.9 Occupant Design Effects by Design Group and Proportion

true confidence intervals for any proportion representing the total area will be much larger than those calculated here.

4.4.4 Summary. The estimated variances and the design effects calculated indicate that some of the proportions have high variances due to the sample design associated with them. The vehicle and occupant level proportions are generally immune to very large clustering effects and for some of them, in particular the injury severity variables, the variances are lower than they would have been from a simple random sample. There the gains from the pre-sampling stratification can be seen. However the variance associated with some accident-level proportions is large. The 95% confidence intervals may be two or three times as large as they would have been from a simple random sample.

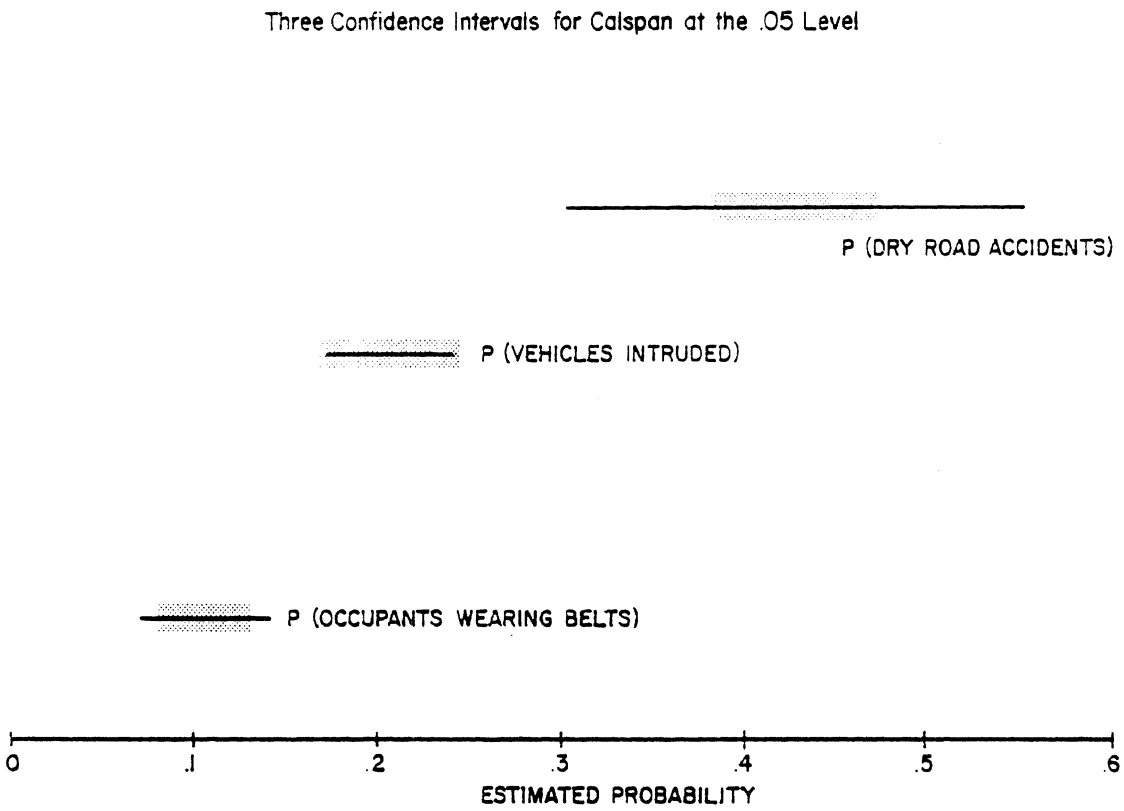

FIGURE 4.10 95% Confidence Intervals for Selected Accident Proportions for Calspan

4.5 Missing Data Analysis

This subsection summarizes the analytical results and problems associated with missing data in the NCSS data. The discussion centers on the key variables, OAIS and Delta V. Section 4.5.1 looks at missing data rates for OAIS and Delta V. The analysis of OAIS is an occupant level analysis and is relatively straightforward. The analysis of Delta V is more complicated. Delta V is a vehicle variable and is only defined for vehicles involved in certain collisions. This makes the calculation of missing data rates and adjusted distributions much more complicated. The bivariate distribution of OAIS and Delta V is done with occupant-level data. In the bivariate case the procedures are similar but Delta V brings along with it its complexities.

Missing data were analyzed separately for fatal and non-fatal occupants. The fatalities were chosen with certainty and were relatively few in number. A complete review of all fatalities was done. A brief description of this subset is given in Section 4.5.2. The three distributions discussed above are then examined for the fatal occupants and those results are presented in Section 4.5.3. The non-fatal occupants with missing data were sampled to obtain 388 non-fatal occupants with missing data on Delta V or OAIS. This sample design is described in Section 4.5.4. For the non-fatals the OAIS distribution was the only distribution studied in depth and this is described in Section 4.5.5. For all occupants, bivariate distributions involving OAIS were calculated to see what effect missing data adjustments had on the cell proportions and marginal distributions. These distributions are presented in Section 4.5.6. Finally the main results from these analyses are summarized in Section 4.5.7.

4.5.1 Extent of Missing Data. There were only a small number of variables in the NCSS data that had substantial missing data. Unfortunately, among these variables were the key variables needed in the modelling task of this project. Delta V, the only variable recorded to measure crash severity, and the CDC variables had substantial missing data. All of the variables that measured injury severity (since information had to come from an official medical source) had missing data in a large proportion of the cases.

Table 4.12 shows the missing data rates for OAIS and Delta V in Phase 1 and Phase 2 for all case vehicle occupants in the NCSS file. The rate at which Delta V and OAIS are both missing is about 15% (unweighted). Since most data analysis packages require data on all variables, 63.6% of the occupants will be excluded from the analyses in Phase 1 and 68.2% in Phase 2. There is a missing data rate of about 30% for OAIS.

TABLE 4.12

Missing Data Rates for Delta V and OAIS
Case Vehicle Occupants

Category		Weighted Distribution (In percent)		Unweighted Distribution (In percent)	
OAIS	DELTA V	Phase 1*	Phase 2	Phase 1*	Phase 2
Miss	Miss	16.54	16.39	15.14	15.00
Miss	Not Miss	16.42	12.05	17.10	12.44
Not Miss	Miss	34.65	43.69	31.18	40.73
Not Miss	Not Miss	32.38	27.87	36.38	31.83
Total		62439	49668	14652	12111

*The official version of the Phase 1 data.

In considering the missing data rate for Delta V, the rate should actually be calculated based only on those vehicles (or occupants) in collisions where Delta V was appropriate to compute. There is no vehicle variable in the NCSS file that indicates for each vehicle whether it would have been appropriate to calculate a Delta V. There is an accident level variable (Crash Reconstruction) which indicates for the accident whether Delta V had been computed. Table 4.13 is similar to Table 4.12 but is restricted to those case vehicle occupants in accidents where a Delta V was calculated. The rate at which data is

missing on both variables is now less than 1% (unweighted) since this now includes only accidents where Delta V was calculated. With this subset of occupants approximately 33% of the occupants will be lost due to missing data on OAIS when doing model estimation.

TABLE 4.13

Missing Data Rates for Delta V and OAIS
Case Vehicle Occupants in an Accident where Delta V was Calculated

Category		Weighted Distribution (In percent)		Unweighted Distribution (In percent)	
OAIS	DELTA V	Phase 1*	Phase 2	Phase 1*	Phase 2
Miss	Miss	1.22	.78	.98	.87
Miss	Not Miss	33.09	25.69	30.89	27.19
Not Miss	Miss	2.14	1.68	2.02	2.35
Not Miss	Not Miss	63.45	71.90	66.11	69.59
Total		31250	23422	8089	5535

*The official version of the Phase 1 data.

To look more closely at the missing data on Delta V for case vehicles, the rate of missing data on Delta V within each category of General Area of Damage for all case vehicles was calculated. Damage in the front, right, back, and left might be a good indirect measure of those situations where the calculation of Delta V might be applicable. Within these categories only approximately 60-72% have a Delta V calculated. All vehicles missing the General Area of Damage had no Delta V calculated. Top and Undercarriage damage would not be applicable for Delta V computation and the missing data rate is predictably high at 95%. If the calculation of the missing data rate for Delta V is based on the four general areas of damage, front, right,

left and rear, 32.4% of these vehicles in Phase 1 and 41.9% of these vehicles in Phase 2 had Delta V missing.

TABLE 4.14

Missing Data Rates for Delta V Within
Types of General Area of Damage
Case Vehicles

General Area of Damage	Delta V			
	Phase 1*		Phase 2	
	Total	Percent Missing	Total	Percent Missing
Missing	1378	99.1	1341	99.6
Front	4532	30.3	3621	40.0
Right	1019	35.2	750	43.3
Back	358	28.6	247	38.9
Left	1014	40.0	791	49.7
Top	322	95.3	355	96.1
Undercarriage	85	89.4	46	97.8

*Official version of the Phase 1 data.

4.5.2 Fatal Supplemental Data. In an early version of the Phase 1 data all fatal cases in the NCSS file were chosen to be further evaluated. This evaluation involved obtaining the hard copy cases and coding as much information about occupant injury, vehicle damage, and crash severity as was possible. This investigation involved 333 fatal occupants. Each fatal occupant was recoded independently from the NCSS codes if present.

More specifically, the information recorded included up to 12 OIC's and an OAIS for each fatality. For each of the injuries and the OAIS the assurance level associated with what was coded was indicated by the

coder and the source in the hard copy that was used to get this information. For each body region the number of injuries received in that body region was recorded. For each of the twelve OIC's the contact that produced the injury was recorded along with the assurance factor and the source of information for each contact recorded. For each vehicle that was involved in the collision that produced the fatality, CDC's were coded and a Delta V was calculated with information that was available in the hard copy. As with the injuries, an assurance factor and an information source was given for each Delta V. The coding instructions for coding OAIS and Delta V instructed the coder to use the unknown code as sparingly as possible. The object of going back over the hard copy cases was to use all the information available in the hard copy and to use experienced subjective judgement to "guess" at values for missing data on OAIS and Delta V. All of the injury coding for these fatals where done by a single coder so the assurance factor is a direct measure of the relative reliability of the OAIS variable. Another coder was responsible for all the crash severity coding.

The option of coding up to twelve OIC's was chosen to allow detailed and complete injury information on fatalities. The average number of injuries to a occupant who received a fatal injury in this file is 7.8. The minimum number of injuries is 1 and the maximum is 37. The median number of injuries is 5. These statistics were developed from a count of the number of injuries to each body region.

4.5.3 Fatal Distributions Adjusted for Missing Data. In this subsection the univariate distributions for OAIS and Delta V are compared for two classes of fatally injured occupants: those with coded data and those with missing data. The data used to obtain the distributions for fatals with missing data was obtained from the supplement described in Section 4.5.2. Once these marginal distributions are considered the bivariate distributions of OAIS and Delta V are compared for these two subsets of the NCSS fatal data.

4.5.3.1 Univariate Distribution of OAIS Adjusted for Missing Data. In the final version of the Phase 1 data of NCSS there were 500

fatalities. Of the 500 fatalities, there were 327[43] fatalities reviewed by HSRI. Of these 327, there were 213 fatal occupants that did not have an OAIS.

There are two ways to modify the distribution for the 500 fatalities to incorporate this additional information:

1. For any fatal occupant missing OAIS use the HSRI recoded value and the original NCSS value for all others.

2. Use a compromise value for those fatal occupants that have both a NCSS OAIS and an HSRI OAIS. For fatal occupants missing OAIS, use the value obtained by the HSRI review.

This subsection, and the following subsections, will only consider the first method mentioned above. The nature of the coding at HSRI of OAIS allows examination of modifications to two versions of the OAIS distribution, the usual distribution of OAIS and the distribution of NEWOAIS3. Table 4.15 shows these distributions for OAIS. The first distribution is the OAIS distribution obtained from the Official Phase 1 Version of the NCSS data. The second distribution is the distribution of all the data recoded by HSRI. The third distribution is the distribution for the missing data in NCSS based on the HSRI recode. The fourth distribution is the adjusted distribution of OAIS incorporating the data available in the NCSS file and the recoded information from the HSRI recode of the supplemental data. Table 4.16 shows the same distributions for NEWOAIS3.

The addition of missing data information to the distribution modifies the OAIS distribution. The largest change in the distribution of the OAIS appears in the proportion of fatalities with an fatalities with an OAIS of 5 before and after the missing data adjustment. In the distribution of NEWOAIS3 the shift is from "injured, severity unknown" to injury at OAIS level 3 or more.

At best, looking at the NEWOAIS3 distribution there is still a 7% missing data rate on the severity of the injury of the fatality. The

---

[43]Some cases coded fatal in the preliminary version of the Phase 1 data were subsequently edited and were not present in the Official Version of the Phase 1 file that this analysis is based on.

TABLE 4.15

Distribution of OAIS
Missing Data Adjustments
Fatal Occupants

Distribution	OAIS 1	3	4	5	6	Injured Severity Unknown	No Information	Total
NCSS	1 .4%	6 2.4%	21 8.4%	110 44%	112 44.8%	250	0	250 100%
HSRI Coded Supplemental Data	1 0.4%	5 1.9%	6 2.3%	39 14.9%	210 80.2%	72	166	334 100%
NCSS Missing (HSRI Recoded)	1 7%	2 1.4%	1 .7%	8 5.4%	136 91.9%	0	186	148 100%
Adjustment for Missing Data	2 .4%	8 1.6%	22 4.4%	118 23.5%	248 49.6%	102 20.4%	0	500 100%

TABLE 4.16

Distribution of NEWOAIS3
Missing Data Adjustments
Fatal Occupants

Distribution	OAIS 0-2	3-6	Injured Severity Unknown	No Information	Total
NCSS	7 2.8%	243 97.2%	250	0	500 100%
HSRI Recoded Supplemental Data	7 2.1%	326 97.9%	1	166	334 100%
NCSS Missing (HSRI Recoded)	4 1.9%	209 98.1%	0	102	213 100%
Adjustment for Missing Data	11 2.2%	452 90.4%	37 7.4%	0	500 100%

situation with OAIS is worse at a 20% missing data rate. This difference in missing data rate is due to coding conventions that were allowed when coding OAIS in the HSRI project. If necessary a choice could be made as to whether the case fell into the OAIS 0-2 category or not, so that a bit more information is available for NEWOAIS3. Table 4.17 compares the injury codes assigned to these occupants.

TABLE 4.17

OAIS Recode Comparison for
Fatal Occupants

HSRI OAIS	NCSS OAIS					Total
	3	4	5	6	Injured Severity Unknown	
1					1	1
3			1		2	3
4	1	3			1	5
5		1	27	3	8	39
6		5	14	53	136	208
Injured Severity Unknown	2	1	2	1	65	71
Total	3	10	44	57	213	327

From the table it can be seen that the largest discrepancy comes from 14 cases coded 5 in NCSS and 6 at HSRI. This may suggest slightly different methods of coding OAIS at the two locations and may serve as a caution in reviewing some of these results.

It should be noted here that there is information for the 65 cases shown in the table as being coded "injured, severity unknown" in NCSS obtained from in the HSRI review. Here, 1 of the 65 was thought to have

an OAIS less than 4 and the other 64 fatals were thought to have an OAIS greater than 3. This information will be used in considering the distribution of NEWOAIS3.

4.5.3.2 Univariate Distribution of Delta V Adjusted for Missing Data. The distribution of Delta V for vehicles in which at least one fatality occurred are discussed in this subsection. There are 426 vehicles involving the 500 fatal occupants that occurred in Phase 1. HSRI recalculated a Delta V for a subset of these 426 vehicles. The 213 vehicles that were recoded contained vehicles that could have a valid Delta V as well as those where a Delta V computation was not applicable.

Delta V presents a situation that is relatively uncommon. There are collisions where Delta V cannot be calculated. The Delta V algorithm requires certain hypotheses (e.g. planar crashes) and makes sense only when these assumptions are satisfied. It is necessary to take this into account when calculating the missing data rate for Delta V and making comparisons of various distributions of Delta V. Calculation of missing data rates and distributions for Delta V should involve only those vehicles where Delta V can be calculated. Vehicles involved in collisions where Delta V cannot be calculated will be considered vehicles where Delta V is not applicable.

The discussion here will focus on Delta V only. Of the 426 vehicles involving a fatality there were 28.1% with unknown Delta V. Of these vehicles an additional 20% were such that the calculation of Delta V was not appropriate. (Based on the Crash Reconstruction variable.) Excluding these 20%, 35.2% of the vehicles where Delta V could be calculated and had a Delta V missing.

The HSRI review supplied information for 84 vehicles of the 120 vehicles where Delta V was coded in NCSS as unknown. For these 84 vehicles it was thought calculation of Delta V was appropriate. Table 4.18 presents the distribution of Delta V in NCSS, the distribution of Delta V for all cases recoded by HSRI, the distribution of the Delta V for all cases missing in NCSS but coded by HSRI and finally a distribution of Delta V modified to use both sources of data.

TABLE 4.18

Distribution of Delta V
Adjusted for Missing Data
Fatal Occupants

Distribution	Delta V										Sub-total	Not Applicable	Total
	Missing	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	Over 45			
NCSS	120	2	8	10	28	24	34	36	21	13	45	221	85	426
		0.9%	3.6%	4.5%	12.7%	10.9%	15.4%	16.3%	9.5%	5.9%	20.4%	100.0%		
HSRI coded Supplemental data	0	0	8	17	18	26	25	28	25	14	52	213	72	284
		0.0%	3.8%	8.0%	8.5%	12.2%	11.7%	13.1%	11.7%	6.6%	24.4%	100.0%		
NCSS Missing (HSRI Recoded)	36	0	3	3	6	10	7	11	9	5	30	84	0	124
		0.0%	3.6%	3.6%	7.1%	11.9%	8.3%	13.1%	10.7%	5.9%	35.7%	100.0%		
NCSS Adjusted for Missing data	30	2	13	17	36	37	44	49	31	19	78	326	70	426
		0.6%	4.0%	5.2%	11.0%	11.3%	13.5%	15.0%	9.5%	5.8%	23.9%	100.0%		

The last distribution, referred to as "NCSS Adjusted for Missing Data," was obtained by using the value that was recoded at HSRI for any case where Delta V was not present in the NCSS data set. This procedure did change the number of vehicles considered "not applicable." Upon review it seemed appropriate to calculate a Delta V for some cases where the crash reconstruction code indicated calculation of Delta V was inappropriate.

From Table 4.18 it can be seen that:

1) The NCSS distribution of Delta V differs from the "NCSS Missing" distribution. cases. The proportion of cases at the highest level of Delta V appears to be higher in the missing cases than the proportion in the non-missing in NCSS.

2) Comparison of the NCSS distribution and the HSRI supplemental data distribution indicate some differences, but there seems to be no clear pattern.

3) The "Adjusted" distribution when compared to the NCSS distribution has a smaller missing data rate. The modification for missing data increases the proportion of greater than 45 mph cases slightly.

4.5.3.3 Bivariate Distribution of OAIS and Delta V Adjusted for Missing Data. In this subsection the joint distribution of OAIS and Delta V for the fatalities in the NCSS data in Phase 1 (Official version) is adjusted for missing data. Ideally we would like to find the distribution of OAIS and Delta V for only those occupants in a towed vehicle where the collision characteristics were such that a Delta V could have been calculated. Because of the presence of missing data on Delta V the subset of occupants defined by Delta V non-missing may give an incomplete subset of occupants. An occupant was included if either NCSS or HSRI determined that calculation of Delta V was possible for the occupant's vehicle. There were 383 fatalities used in this analysis. Table 4.19 shows the distribution of these fatalities by Delta V source. Approximately one third of the fatalities had a Delta V calculated by both HSRI and NCSS. Another third had a Delta V calculated only by NCSS. (These were vehicles that were not in the

subset reviewed by HSRI). The remaining third had a Delta V coded by only HSRI. The variable used from NCSS to determine where Delta V could be calculated was Crash Reconstruction while HSRI made this judgement directly.

TABLE 4.19

Fatalities
Included in the
Bivariate Distribution

Category	NCSS Valid Delta V	HSRI Valid Delta V	Number of Occupants
1	Yes	Yes	129
2	Yes	No	3
3	Yes	Unk	128
4	No	Yes	22
5	No	No	0
6	No	Unk	0
7	Unk	Yes	101
8	Unk	No	0
9	Unk	Unk	0
Total			383

Table 4.20 presents the NCSS distribution of OAIS by Delta V for these 383 occupants recoded into 6 categories. The categories for Delta V are defined by 1-4, 5-9, 10-14, 15-19, 20-24, and over 25. The percentages under the counts are the row percentages. The percentages and totals in parentheses exclude all missing data.

Table 4.21 is the distribution of OAIS by Delta V categorized the same as above for the 127 fatalities in Table 4.20 with missing Delta V. This distribution is based on the missing data coded by HSRI in the missing data supplement.

Table 4.22 is the NCSS distribution of OAIS and Delta V supplemented with additional information about the missing data. The missing data rate has now been reduced to less than 1%. Both the injury variable, OAIS, and Delta V were adjusted for missing data. This

TABLE 4.20

Distribution of Delta V and OAIS
NCSS Distribution
Fatal Occupants

OAIS	Delta V							Total
	Missing	1-4	5-9	10-14	15-19	20-24	Over 25	
1.	0	0	1 100% (100%)	0	0	0	0	1 (1)
3.	2 50.0%	0	1 25.0% (50.0%)	1 25.0% (50.0%)	0	0	0	4 (2)
4.	5 31.3%	0	3 18.8% (27.3%)	2 12.5% (18.2%)	4 25.0% (36.4%)	1 6.3% (9.0%)	1 6.3% (9.0%)	16 (11)
5.	28 30.4%	3 3.3% (4.7%)	8 8.7% (12.5%)	13 14.1% (20.3%)	23 25.0% (39.1%)	8 8.7% (12.5%)	9 9.8% (14.1%)	92 (64)
6.	31 36.5%	3 3.5% (5.6%)	10 11.8% (18.5%)	12 14.1% (22.2%)	12 14.1% (22.2%)	9 10.6% (16.7%)	8 9.4% (14.8%)	85 (54)
Unknown	61 33.0%	4 2.2% (3.2%)	12 6.5% (9.7%)	31 16.8% (25.0%)	34 18.4% (27.4%)	17 9.2% (13.7%)	26 14.1% (21.0%)	185 (124)
Total	127 33.2%	10 2.6% (3.4%)	35 9.1% (13.7%)	59 15.4% (23.0%)	73 19.1% (28.5%)	35 9.1% (13.7%)	44 11.5% (17.2%)	383 (256)

adjustment procedure simply replaced a missing value (or unknown for OAIS) in NCSS with the HSRI coded value if it exists, otherwise it is left missing.

Table 4.23 presents the marginal distribution of OAIS for these 383 fatalities. These fatalities represent a special subpopulation of the total population; that is, occupants of vehicles where the vehicle dynamics were such that a Delta V could be calculated.

TABLE 4.21

Distribution of Delta V and OAIS
HSRI Coded Supplemental Data
Fatal Occupants

OAIS	Delta V							Total
	Missing	1-4	5-9	10-14	15-19	20-24	Over 25	
Missing	1 100%	0	0	0	0	0	0	1
4.	0	0	2 50.0%	1 25.0%	0	0	1 25.0%	4
5.	1 4.8%	0	2 9.5%	6 28.6%	2 9.5%	5 23.8%	5 23.8%	21
6.	1 1.2%	4 5.0%	8 10.0%	11 13.8%	26 32.5%	10 12.5%	20 25%	80
Unk	0	0	2 9.5%	7 33.3%	5 23.8%	2 9.5%	5 23.8%	21
Total	3 2.4%	4 3.1%	14 11.0%	25 19.7%	33 26.0%	17 13.4%	31 24.4%	127

From these tables the following observations can be made:

1) The adjusted row percentages in Table 4.22 are close to the NCSS percentages ignoring the missing data in Table 4.20. The percentage in the Delta V category "over 25" is consistently lower for all AIS levels in the NCSS data (excluding missing data). The NCSS percentage (excluding missing data) is lower for some of the AIS categories when Delta V is between 20 and 24. There appears to be no other clearly consistent trends.

2) Comparing the modified distribution in Table 4.23 with the NCSS distribution (excluding missing data) from Table 4.20 it can be seen that the NCSS distribution (excluding missing data) underestimates the percentage of fatalities with an OAIS of 6 and overestimates the number of fatalities with an OAIS of 5.

TABLE 4.22

Distribution of OAIS and Delta V
Adjusted for Missing Data
Fatal Occupants

Missing	Delta V							Total
	Missing	1-4	5-9	10-14	15-19	20-24	Over 25	
Missing	1	1	2	3	3	2	7	19
1.	0	0	1 100%	0	0	0	0	1
3.	0	0	2 33.3%	2 33.3%	1 16.7%	0	1 16.7%	6
4.	0	1 5.9%	5 29.4%	3 17.6%	4 23.5%	1 5.9%	3 17.6%	17
5.	2 2.0%	4 4.0%	12 12.1%	24 24.2%	31 31.3%	14 14.1%	12 12.1%	99
6.	0	7 3.6%	24 12.4%	38 19.6%	54 27.8%	27 13.9%	44 22.7%	194
Unk	0	1 2.1%	3 6.4%	14 29.8%	13 27.7%	8 17.0%	8 17.0%	47
Total	3 0.8%	14 3.7%	49 12.8%%	84 21.9%	106 27.7%	52 13.6%	75 19.6%	383

It should be noted here that the coding rules of OAIS were different for the NCSS and HSRI coders. The objective was to get information not originally coded to see how the NCSS distribution might possibly change if all the data had been obtained originally.

TABLE 4.23

Marginal Distribution of
OAIS for Occupants in
Vehicles Where Delta V Could Be Calculated

	Missing	1.	3.	4.	5.	6.	Total
NCSS Distribution (including M.D.)	185 48.3%	1 .3%	4 1.0%	16 4.2%	92 24.0%	85 22.2%	383
NCSS Distribution (excluding M.D.)		1 8%	2 1.5%	11 8.3%	64 48.5%	54 40.9%	132
Missing Data HSRI Coded	22 17.3%			4 3.1%	21 16.5%	80 6.3%	127
Modified Distribution	66 17.2%	1 .3%	6 1.6%	17 4.4%	99 25.8%	194 50.7%	383

4.5.4 Non-fatal Supplemental Data In the NCSS data, a large percentage of non-fatal occupants investigated have missing data on either Delta V or OAIS or both. This subsection describes the procedure by which a subsample of these occupants was obtained. Each case was examined to get values for Delta V and OAIS or information about why the information was unobtainable. This information will be used to investigate missing data biases in NCSS.

A sample was taken of all occupants in the Phase 1 NCSS occupant file (N=14491 case vehicle occupants) who were not fatally injured and for which either or both of the values for Delta V and OAIS were missing. There were 9035 such occupants, distributed as shown in Table 4.24 where the percent (shown in parentheses) is the percent of all "missing data" cases. Note that of the 9035 cases in all which have either OAIS, Delta V, or both, missing 21.8% of these cases were missing both elements.

TABLE 4.24

Frequency of Missing Data
on OAIS and Delta V
Case Vehicle Occupants

Delta V	OAIS		Total
	Missing	Not Missing	
Missing	2186	4514	6700
Not Missing	2335		2335
Total	4521	4514	9035

Since partial information on Delta V and OAIS is available for some occupants, the frame was divided into three groups:

Group 1 - Occupants where Delta V and OAIS were missing

Group 2 - Occupants where only Delta V was missing

Group 3 - Occupants where only OAIS was missing

The sampling design for the sample of Groups 2 and 3 used the information about Delta V or OAIS, when available, in the stratification of the occupants.

Stratification for the occupants in Group 1 was based on team, age, and NCSS class. Age was coded into two groups: less than or equal to 35 and greater than 35. NCSS class was recoded into four groups: hospitalized (NCSS class code 4), transported (NCSS class code 5), unknown treatment (NCSS class codes 6,7,9) and no treatment (NCSS class code 8). The distribution of occupants in Group 1 by age and NCSS class is given in the Table 4.25 below.

TABLE 4.25

Stratification of Occupants in Group 1
by NCSS Class and Age

Age	NCSS Class				
	Hospitalized	Transported	Unknown Treatment	No Treatment	All
Less than 36 .	201	453	443	343	1530
Greater than 35	107	154	265	130	656
All	398	607	708	473	2186

This final stratification, by team, age, and NCSS class, had 56 strata (7x2x4), the largest stratum with 139 occupants and the smallest with 2 occupants.

Group 2 was, in addition to team and age, stratified by OAIS since all occupants have an OAIS coded. Age was recoded as defined above and OAIS was recoded into four groups: uninjured (OAIS code 0), minor (OAIS code 1), moderate (OAIS code 2) and more than moderate (OAIS codes 3-5). The distribution of occupants is given in Table 4.26. The final stratification by OAIS, age, and team, had 56 strata (4x2x7), the largest stratum with 380 occupants and the smallest with 4 occupants.

TABLE 4.26

Stratification of Occupants in Group 2 by Age and OAIS

Age	Uninjured	Minor	Moderate	Severe	Total
Less than 36 .	1598	965	338	283	3184
Greater than 35	665	362	143	160	1330
Total	2263	1327	481	443	4514

Finally in Group 3 the stratification variables were age, team, Delta V, and General Area of Damage. Age was recoded as previously stated, General Area of Damage was recoded into 2 groups (front versus elsewhere), and Delta V was recoded into 3 groups (less than 10, 10-19, and 20+). The distribution of occupants in Group 3 is given below in Table 4.27. The final stratification by team, age, Delta V, and General Area of Damage had 84 strata (7x2x3x2), the largest stratum with 203 occupants. There was 1 stratum which was empty and two strata with only one occupant.

TABLE 4.27

Stratification of Group 3 by Age, Delta V, and General Area of Damage

Delta V	General Area of Damage			
	Age less than 36		Age greater than 35	
	Front	Other	Front	Other
Less than 10	254	179	140	99
10-19 . . .	563	315	215	149
Over 20 . .	251	86	64	20
All	1068	580	419	268

The sampling, for all three groups, was a simple random sample without replacement of two occupants within each stratum. For all strata with less than 3 occupants a complete census of occupants in those strata was made. A total of 388 occupants are in the sample.

4.5.5 Non-fatal Distributions Adjusted for Missing Data. The previous subsection described the sampling plan by which a sample of non-fatal occupants missing either Delta V or OAIS was chosen. In this subsection the distribution of OAIS for occupants with missing data based on the supplemental data on non-fatals is examined. It should be noted that in this subsection "missing data" refers to missing data on either OAIS or Delta V.

For this discussion the NCSS sample can be separated into

S_1: All non-fatal occupants with complete data on OAIS and Delta V, and

S_2: All non-fatal occupants with missing data on either OAIS or Delta V or both.

This discussion will focus on differences between the occupants in S_1 and S_2.

There are three distributions presented in Table 4.28. The unweighted distribution from NCSS for occupants with no missing data is the distribution of S_1 in Phases 1 and 2. The distribution from the supplemental data is the estimated distribution for the occupants in the sample that belong to S_2. Associated with the estimated distribution for S_2 are the standard errors for the proportions.

From Table 4.28 it can be seen that the distributions across categories differ for the occupants representing S_1 and S_2. The estimated distribution of S_2 has a higher proportion with a minor injury. The distribution of S_1 has a higher proportion of occupants with no injury. Looking at the cumulative proportion of injury with severity less than OAIS 3 makes the distributions more comparable. The proportion of injuries less than OAIS 3 is 87.7% for S_1 From the estimated distribution of S_2 it can be seen that 92.70% of these occupants have injuries less than OAIS 3. The most significant difference between the distribution of S_1 and the estimated distribution

TABLE 4.28

DISTRIBUTION OF OAIS

Non-fatal Occupants

OAIS Category	Unweighted[a] Distribution (In percent)		Supplemental Data[b] (Weighted) (In percent)	
	Phase 1[c]	Phase 2	Estimated Proportion	Standard Error
0.No Injury	37.5	31.4	22.48	2.21
1.Minor	36.8	37.0	42.8	2.55
2.Moderate	13.4	16.6	12.3	2.24
3.Severe	8.6	10.5	6.72	.92
4.Serious	2.6	3.5	1.37	0.4
5.Critical	1.12	0.8	0.33	0.24
9.Unknown			1.60	.38
0.-1.			9.86	2.49
0.-2.			3.26	1.04
Sample Size	14153	11587		

[a]Includes all non-fatal occupants with no missing data on OAIS.

[b]Estimate of a distribution for non-fatal occupants with missing data on either OAIS or Delta V missing.

[c]The official version of the Phase 1 data.

of S_2 is the increase in the percentage of minor injury. The reason for this is inherent in the NCSS coding instructions for OAIS. OAIS can only be coded if there is an official medical source for the injury information. Therefore minor injuries are most likely to be coded missing data due to the lack of medical records. When looking at severe

versus non-severe injury the distribution of this dichotomy does not appear to be significantly different.

In order to get an estimate of the distribution for the sample a weighted average of the two estimated distributions in Table 4.28 can be used. The weights are the proportions of the sample in S_1 and S_2. Note that this is not an estimate for the aggregate of the NCSS areas but only for the sample. In order to approximate a distribution for the aggregate it would be necessary to know the proportions in the aggregate that belong to each subpopulation.

4.5.6 Bivariate OAIS Distributions Adjusted for Missing Data. As in the preceding subsection the discussion here will focus on differences between the occupants (fatal and non-fatal) completely coded on OAIS and Delta V, S_1, and the occupants (fatal and non-fatal) with missing data for at least one of the two variables, S_2. The objective is to evaluate the differences in the observed and unobserved occupants.

Bivariate distributions involving OAIS are investigated in this subsection. When comparing bivariate distributions for S_1 and S_2 the change in the association between the variables in important. When missing data is excluded any association represented in the observed table is used as an estimate for the population. The presence of information about the missing data could change the association in the table substantially.

Four bivariate tables were investigated. The variables related to OAIS were seat position, sex, urbanization, and restraint usage. The estimates in this subsection were obtained using information on missing data for both fatal and non-fatal occupants. A direct substitution was made for all fatals where there was supplemental information available. The frequencies in the contingency table for S_2 were obtained using the weights resulting from the sample design described in Section 4.5.4. The distribution of S_1 is determined from all occupants in the sample with complete information on OAIS and Delta V. Cell proportions were then calculated to allow comparisons to be made between the distribution of S_1 and the estimated distribution of S_2.

Tables 4.29 and 4.30 show the results for the bivariate distributions of OAIS with Restraint Usage and Urbanization. The bivariate tables of OAIS and urbanization showed the most difference of all of the tables calculated. Bivariate distributions of OAIS and Restraint Usage was typical of the distributions calculated. In the tables there are two distributions, the distribution (unweighted) from the NCSS for complete data cases and the distribution estimated for the missing data occupants.

Again it can be seen that the marginal distribution of OAIS is different for those occupants not observed. The major change is the increase of the probability of minor injury. The marginal distribution of urbanization changes slightly and the marginal distribution of Restraint usage stays fairly constant. Notable is the changes of the proportions within each cell. In both tables the cell associated with lower levels of OAIS change noticeably over levels of the other variable.

TABLE 4.29

Distributions of
the Degree of Urbanization and OAIS

OAIS	NCSS Distribution Non-missing Data Unweighted percents			Estimated Distribution for Missing Data Weighted Percents		
	Urban	Rural	Total	Urban	Rural	Total
0	9.46	32.11	41.57	6.52	22.44	28.96
1	9.91	21.65	31.56	10.52	28.54	39.06
2	4.29	7.10	11.41	3.76	7.63	11.39
3	3.44	4.22	7.66	3.43	3.94	7.37
4	1.05	1.32	2.37	0.95	0.93	1.88
5	0.96	1.05	2.11	0.66	0.72	1.38
6	1.55	0.85	2.40	1.07	0.59	1.66
Injured	0.71	0.28	.99	1.79	6.50	8.29
Total	31.40	68.60		28.71	71.29	

TABLE 4.30

Distributions of
Restraint Usage by OAIS
Adjusted for Missing Data

OAIS	NCSS Distribution Non-missing Data Unweighted Percents				Estimated Distribution for Missing Data Weighted Percents			
	No	Yes	Unknown	Total	No	Yes	Unknown	Total
0	29.47	5.42	6.67	41.56	20.57	3.80	4.59	28.96
1	25.76	3.26	2.55	31.56	30.58	5.53	2.95	39.06
2	9.66	0.96	0.79	11.41	9.30	1.45	0.65	11.39
3	6.52	0.65	0.49	7.66	6.06	0.82	0.48	7.37
4	2.08	0.16	0.14	2.37	1.64	0.11	0.13	1.88
5	1.63	0.21	0.17	2.11	1.12	0.15	0.12	1.38
6	1.85	0.29	0.27	2.40	1.27	0.20	0.19	1.66
Injured	0.84	0.09	0.06	0.99	5.72	0.45	2.12	8.29
Total	77.81	11.04	11.14		76.27	12.50	11.22	

4.5.7 Summary The analysis presented in Section 4.5 on missing data adjustments is just the beginning of serious work on the problem of missing data. The analysis presented restricts itself to very simple methods of adjustment and concentrates on differences between the "missing" and "non-missing" segments of the sample observed.

The two key NCSS variables, OAIS and Delta V, have exceptionally high missing data rates. The OAIS missing data rates for all occupants is approximately 30%. For Delta V the missing data rate, based on the General Area of Damage, is approximately 34% in Phase 1 and 41.7% in Phase 2. When the joint distribution of Delta V and OAIS is considered for all occupants, 72% in Phase 1 and 68.2% in Phase 2 one or more of these two variables are missing.

The results from the adjustment of distributions involving fatalities show that the proportion of occupants with maximum injury (OAIS-6) is higher after adjustment and the proportion of fatalities at higher Delta V is slightly higher after adjustment. The bivariate distribution is consistent with these results. The distribution appears shifted slightly to higher OAIS levels.

Examination of the non-fatal missing data occupants shows a significant increase in the probability of a minor injury (OAIS 1). The bivariate distributions for all occupants involving OAIS change both marginal distributions, the distribution of OAIS changing the most obviously. Cell proportions also show differences between missing and non-missing data.

From these analyses it appears that the missing data population has a different distribution for OAIS and Delta V than the non-missing population has. The adjusted distributions calculated indicate that for some categories the proportion represented by the NCSS data (excluding missing data) are reasonable estimates. But there are some categories that the NCSS proportions will over-estimate or under-estimate without adjustments for missing data. It appears that the dichotomous variable based on OAIS is less sensitive to missing data.

Based on these analyses it is recommended that an imputation procedure procedure be developed for the NCSS data. The imputation

procedure can be chosen in such a way that the distributions of OAIS and Delta V are modified in a manner consistent with the distribution of the non-missing data on OAIS and Delta V. These imputation procedures are alternatives to just excluding missing data and can produce adjusted distributions more representative of the population. Choice of which particular imputation method is best is a subject for further investigation.

4.6 National Projections

In this subsection nationally representative statistics using the NCSS data are discussed and some national projections are given for selected NCSS variables. The procedure to generate national projections is demonstrated for simple accident statistics and for more complicated contingency tables. Methods for adjusting these national projections for missing data are developed. Application of this methodology will depend on the particular missing data adjustment technique selected. Finally these national projection are evaluated to see how sensitive they are to assumptions made in the development of the estimation procedure.

Some of the national projections developed in Section 4.6.1 are compared with estimates obtained using other methods. These methods use an inflation factor applied to the aggregate NCSS estimate. There is no way to decide which method yields a better nationally representative estimate. A choice between the methods depends on the reasonableness of the assumptions underlying method.

4.6.1 NCSS National Projections. The NCSS design called for a sample of accidents (police reported accidents involving at least one towed vehicle) within each area chosen. Even though the sample design differed between areas, the design was such that the sample design within each area ensured a proper sample within each county. For the purpose of producing national projections, counties were chosen to represent the unit of observation. This choice was made because of the size of the geographical unit, the well-defined structure of the county, and the availability of demographic data for each county. Another possible choice would have been the areas themselves. Since this choice would have reduced the number of observations on which the regression model could have been developed, it was not considered further.

The seven areas chosen for NCSS are composed of 41 counties and three partial counties. The three partial counties are defined by Miami city, Erie county (omitting the city of Buffalo), and three police districts in Los Angeles city. In the following analysis 43 observations are used. These are the 41 counties, Miami city, and Erie county (omitting Buffalo). The Los Angeles data were not used in this

analysis because very few demographic data were available for those three police districts in Los Angeles. It is advisable if possible to use all NCSS data from all seven areas when making NCSS projections. in the development. It was possible to calculate all the demographic statistics for the two partial counties because data were available for Erie county and also for Miami and Buffalo so that these two areas were included. Similar data were not available for the three Los Angeles police districts so they were excluded.

Three traffic accident totals were chosen to be estimated. These statistics are defined by:

ACCIDENTS = the number of accidents which involve a towed passenger car (NCSS accidents),

VEHICLES = the number of towed passenger cars in NCSS accidents for which Delta V could be calculated (this excludes vehicles in rollovers, sideswipes, underrides, overrides, accidents with vaulting, and collisions with yielding fixed objects),

OCCUPANTS = the number of occupants in towed passenger cars where Delta V could be calculated.

In the NCSS data set the information on the above quantities is essentially complete[44]; that is, every accident investigated, along with all towed vehicles and their occupants, is represented in the file. It is assumed that all eligible accident reports were sampled and no accidents were "lost" in the process of obtaining accident reports. Violation of this assumption will clearly affect the accuracy of the national projections, the end result being that any national projection using the NCSS data will likely underestimate the actual national total.

Within each of the 43 observed counties an unbiased estimate for the county total of ACCIDENTS, VEHICLES, and OCCUPANTS was produced. The county estimates for ACCIDENTS, VEHICLES, and OCCUPANTS are given by the following formula:

---

[44]The version of the NCSS data used in this analysis was a preliminary version and was missing 55 accidents.

$$(4\text{-}33) \qquad t_k = \sum_{i=1}^{3} t_{ik}/p_i$$

where

t_k is the total for the k^{th} county,

t_{ik} is the total for the k^{th} county represented in the i^{th} stratum, and

p_i is the probability of selection for the i^{th} stratum.

These statistics were then merged with a data file which contains demographic information for each county.

For this analysis, twenty demographic variables were chosen out of the County and City Data Book[45] that were thought to be related to these traffic accident statistics. Correlations were done initially to narrow the search to demographic variables where a linear model fit well. Table 4.31 gives the correlations for five demographic variables investigated to show the range of correlations possible.

TABLE 4.31

Correlation Analysis

Variable	Accidents	Vehicles	Occupants
AUTO DEALERS[a]	.9935	.9860	.9880
GAS STATIONS[b]	.9689	.9511	.9556
LAND AREA[c]	.1206	.0993	.1062
POPULATION[d]	.9463	.9251	.9315
DENSITY[e]	.5476	.6028	.5904

[a]Retail sales by automotive dealers
[b]Retail sales by gasoline service stations.
[c]Area in square miles.
[d]The 1975 population total estimate.
[e]The ratio of 1975 population to land area.

[45]U. S. Bureau of the Census, County and City Data Book, 1977: A Statistical Abstract Supplement (Washington, D.C.: Government Printing Office, 1978).

The three demographic variables, AUTODEALERS, GAS STATIONS, and POPULATION were chosen to investigate further. Scatter plots were obtained to check for nonlinearities in the data. Because of its high correlation with all three accident statistics of interest, AUTODEALERS was chosen to use to develop the model from which the value for the unobserved counties would be predicted. (Accidents are plotted against retail sales by autodealer in Figure 4.11. Plots for vehicles and occupants look similar.) Linear and quadratic models using AUTODEALERS were fit to each of the three accident variables. Table 4.32 presents the results from these regression analyses. The linear fit explained over 97% of the variability in each of the three accident statistics. The addition of the quadratic term did not seem justified. The final predictive equations to be used to predict for the three accident statistics in counties <u>not</u> observed are given by the following equations:

$$\begin{aligned} \text{ACCIDENTS} &= 35.298 + .026396\ \text{AUTODEALERS} \\ \text{VEHICLES} &= 16.580 + .021520\ \text{AUTODEALERS} \\ \text{OCCUPANTS} &= 6.770 + .033348\ \text{AUTODEALERS} \end{aligned} \tag{4-34}$$

TABLE 4.32

Regression Analysis Summary

Variables	R^2		Standard Error	
	Linear	Quadratic	Linear	Quadratic
ACCIDENTS	.987	.988	244	238
VEHICLES	.972	.973	293	294
OCCUPANTS	.976	.976	420	424

The national projection is composed of the sum of the estimates for each of the observed counties and the sum of the predictions, given by the equations above, for each of the unobserved counties. Table 4.33 shows the estimate for observed counties (excluding the two partial

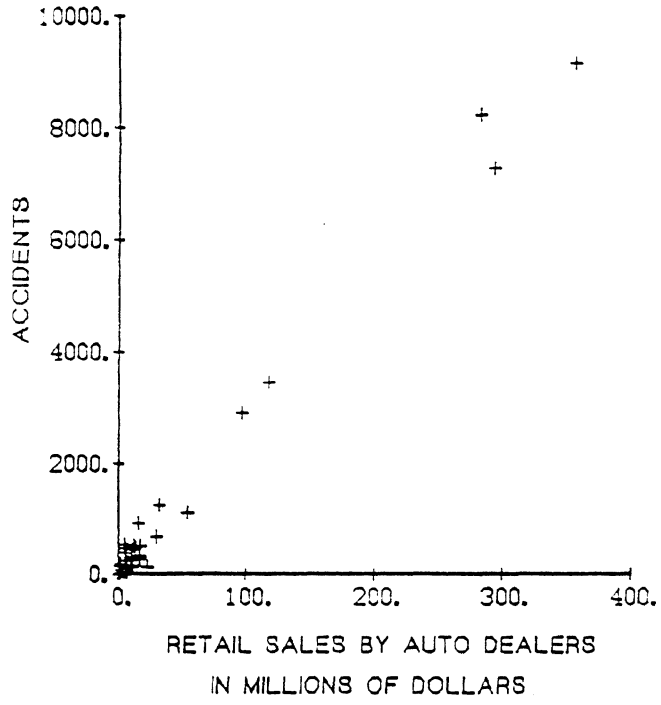

FIGURE 4.11 A Plot of ACCIDENTS By Auto Dealer Retail Sales

counties), the prediction for all unobserved counties (including Erie and Dade counties), and the national projection.

TABLE 4.33

National Projections of the
Accident Statistics
January 1977 through March 1978

Variables	Projected Total for Unobserved Counties (N=3071)	Estimated Total for Observed Counties (N=41)	National Projections
ACCIDENTS	2,420,353	26,729	2,447,082
VEHICLES	1,833,962	19,908	1,853,870
OCCUPANTS	2,900,039	31,754	2,931,793

The selection of the forty-three counties included in NCSS contributes to the bias in the national projection. As seen from Equation 4-31 in Section 4.1, one source of bias in the national projection is $b(N-n)(X-\bar{x}_s)$. For AUTODEALERS the difference $(X-\bar{x}_s)$ is -7360. So part of the bias is given by $-7360b(N-n)$. If the true value of b were known, some adjustment could be made to the national projection, but since b is an unknown parameter, the information about the exact magnitude of this source of bias is incomplete. If δ is the true value for b, then the contribution toward the bias from this term would be about 596,617. The confidence interval for δ is (.02545, .02734). Using this information the bias might be in the range -617,863 to -575,235. If this were the only contribution to the bias the national projection for accidents would be an underestimate of the national total of accidents involving towed vehicles.

The second source of bias comes from the sum of each of the theoretical deviations of the totals from the true population linear model, that is, the sum of $e_i = Y_i - a - bX_i$ $e_i=1,\ldots,n$. There is no direct way to estimate the e_i from the data. If all of the e_i are close to zero then any subset of counties should produce approximately the same regression line and the variability of the projections for different subsets should be small. On the other hand, if all of the e_i are large, then great variability should show up both in the model's fit and the projections. If only a small number of the e_i are large, then the results of the regression modelling and the predictions should be sensitive to whether those counties with large e_i are included in the chosen set. But for any particular subset there is no way to know for sure which situation fits reality.

The method of national projections can be applied to contingency tables. The summary statistics (for the aggregate) in a contingency table can be adjusted to be better representative of the associated contingency table for the population. This is accomplished by calculating a national projection for each frequency count in the contingency table. To demonstrate this technique the national projection for the bivariate distribution of case vehicle occupants by the NCSS classification and the general area of damage from the CDC was

used. The NCSS classification was re-categorized into four groups, fatal, hospitalized, treatment and no treatment. The definition of these categories is give in Table 4.34. Occupants are restricted to those whose vehicles had front, right, rear or left damage.

A national projection for this contingency table involves sixteen national projections, one for each cell frequency. For this analysis there are sixteen variables for which a good prediction equation from demographic data must be found. Twenty variables from the census data were chosen and correlations were examined. It was decided to use 1975 population as the independent variable in all sixteen regressions. The 1975 population had high correlations with all variables. Correlations of the sixteen frequencies with 1975 population ranged from .988 to .581. The regressions were done and the results from the regression analysis are presented in Table 4.34. In most of the regressions the estimated intercepts were quite small. But there appeared to be quite a bit of variability in the estimated slopes. The largest estimate for the slope was .0048 and the smallest estimated slope was .0000012. These regression equations were then used to predict each of the sixteen frequencies for all of the counties not chosen in NCSS. These national projections are given in Table 4.35.

One important question to consider is whether the distribution based on the NCSS aggregate frequencies differs from the distribution based on the national projection frequencies. The distribution based on the NCSS aggregate, the distribution based on predictions for the unobserved counties obtained using the regression models and the distribution based on the national projections were compared. These three distributions are given in Tables 4.36 to 4.38. The distribution of the NCSS aggregate does not differ much from the distribution of the national projections. It should be noted that the distribution of the national projection is closer to the distribution based on the predictions for the unobserved counties. This is due to the fact that only 43 of the 3112 counties were observed in NCSS. The national projection is more heavily weighted by the predictions for the unobserved counties.

TABLE 4.34

Regression Analysis Summary
for the National Projections of the
Distribution of Occupants by NCSS Class and General Area of Damage

Dependent Variable	Regression Statistics			
	Estimated Intercept	Estimated Slope	R^2	Standard Error
FRONT-FAT[a]	2.1178	0.35228×10^{-4}	0.90974	1.97
FRONT-HOS[b]	7.7916	0.39593×10^{-3}	0.96837	12.73
FRONT-TRT[c]	20.372	0.14992×10^{-2}	0.87128	102.50
FRONT-NTRT[d]	1.0145	0.47566×10^{-2}	0.97631	131.80
RIGHT-FAT	0.60184	0.19715×10^{-4}	0.81081	1.69
RIGHT-HOS	2.0586	0.95013×10^{-4}	0.87406	6.42
RIGHT-TRT	4.9504	0.32372×10^{-3}	0.77785	30.77
RIGHT-NTRT	5.9745	0.10022×10^{-2}	0.89386	61.41
BACK-FAT	-0.028479	0.12409×10^{-5}	0.41694	0.26
BACK-HOS	0.22552	0.17125×10^{-4}	0.33712	4.27
BACK-TRT	2.2825	0.15416×10^{-3}	0.73424	16.50
BACK-NTRT	0.40072	0.39616×10^{-3}	0.90238	23.178
LEFT-FAT	0.25263	0.20307×10^{-4}	0.78767	1.88
LEFT-HOS	1.9517	0.87846×10^{-4}	0.85888	6.33
LEFT-TRT	2.5514	0.36577×10^{-3}	0.72496	40.08
LEFT-NTRT	-3.7712	0.13053×10^{-2}	0.96711	42.82

[a]NCSS classification coded 1, 2, or 3.

[b]NCSS classification coded 4.

[c]NCSS classification coded 6.

[d]NCSS classification coded 8.

4.6.2 Evaluation of the Model. The effect of specific counties was investigated to analyze how sensitive the national projection is to the choice of a particular county. One of the traffic accident statistics, ACCIDENTS, was chosen for this investigation. This analysis consisted of using 42 counties (one less than the total number available) to produce a national projection. This procedure was done 43 times. Each time the analysis was done a different county was excluded. Varying the particular counties used in fitting the model, on which the predictions for the unobserved counties are based, does change the value

TABLE 4.35

Frequency Distribution
NCSS Classification and General Area of Damage
National Projections
January 1977 to March 1978

General Area of Damage	NCSS Class			
	Fatal	Hospitalized	Treatment/ Not Hosp.	No Treatment
Front	14,046	108,102	380,316	1,010,512
Right	6,052	26,502	83,799	230,856
Back	176	4,310	39,679	85,113
Left	5,088	24,655	85,145	264,688

TABLE 4.36

Distribution of
NCSS Classification and General Area of Damage
NCSS Aggregate
January 1977 to March 1978

General Area of Damage	NCSS Class			
	Fatal	Hospitalized	Treatment/ Not Hosp.	No Treatment
Front	0.0067	0.0500	0.1499	0.4188
Right	0.0030	0.0112	0.0317	0.1031
Back	0.0001	0.0012	0.0152	0.0358
Left	0.0024	0.0105	0.0277	0.1127

of the projection. Figure 4.12 is a histogram of 43 national projections obtained in this investigation and descriptive statistics can be found in Table 4.39.

From the histogram and the descriptive measures it can be seen that the national projections do vary as different counties are excluded from the analysis. There is no way to tell which of these national projections is closer to "true" population total. The minimum and

TABLE 4.37

Distribution of
NCSS Classification and General Area of Damage
Predictions for Unobserved Counties
January 1977 to March 1978

General Area of Damage	NCSS Class			
	Fatal	Hospitalized	Treatment/ Not Hosp.	No Treatment
Front	0.0059	0.0456	0.1606	0.4264
Right	0.0025	0.0112	0.0354	0.0974
Back	0.0001	0.0018	0.0168	0.0359
Left	0.0021	0.0104	0.0360	0.1117

TABLE 4.38

Distribution of
NCSS Classification and General Area of Damage
National Projections
January 1977 to March 1978

General Area of Damage	NCSS Class			
	Fatal	Hospitalized	Treatment/ Not Hosp.	No Treatment
Front	0.0059	0.0456	0.1605	0.4265
Right	0.0026	0.0112	0.0354	0.0974
Back	0.0001	0.0018	0.0167	0.0359
Left	0.0021	0.0104	0.0359	0.1117

maximum national projections presented in Table 4.39 might be thought of as upper and lower bounds for the national projection if only 42 of the counties originally selected were observed. This discussion is not an analysis of the sampling variance of the national projection which must take into account the uncertainly in the estimates of the population totals for each observed county. Rather it is an evaluation of how the choice of observed units might have affected the national projection.

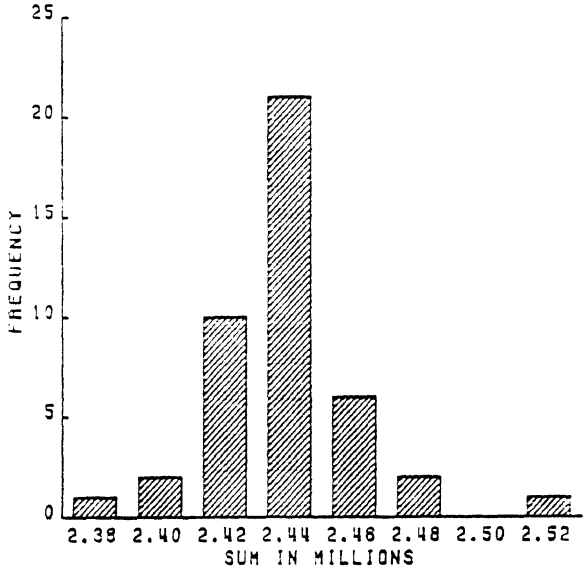

FIGURE 4.12 A Histogram of Possible National Projections Obtained by Varying Counties Included

TABLE 4.39
Descriptive Statistics for the
County Selection Sensitivity Analysis

Parameter	Estimate (N=43)
Mean	2,448,969
Standard Deviation	138,565
Minimum	2,397,501
Maximum	2,527,067

One final question deserves a bit of attention. That is, how important is the choice of the demographic variable that is used in the prediction equation, in this estimation process. A sensitivity analysis was done to investigate this problem. After a thorough review of the County and City Data Book twenty demographic variables were found for

use in this sensitivity analysis. Each of the twenty demographic variables was used to develop national projections for the number of towed vehicles (not restricted to those vehicles where Delta V can be calculated). Table 4.40 gives some descriptive information obtained from this analysis. For the twenty different demographic variables the national projections ranged from approximately 1.86 million to 4.05 million. This information suggests that the choice of the demographic variable can substantially change the projection of the number of towed vehicles.

TABLE 4.40

Descriptive Statistics for the
Demographic Variable Sensitivity Analysis

Parameter	Estimate (N=20)
Mean	2,719,800
Standard Deviation	463,310
Minimum	1,859,000
Maximum	4,047,400

To further investigate this variability, the national projection was plotted as a function of R^2 from the regression producing the prediction equation developed to make the national projection. This plot is shown in Figure 4.13. From this plot it can be seen that the extreme national projection are associated with regressions that have low R^2. The variability of the national projection appears to decrease slightly as R^2 increases, but in general the projections all lie within 2,292,300 and 3,244,200 when demographic variables whose R^2 are greater than .6 are used in developing the national projections. There does not appear to be a "true value" for the number of towed vehicles appearing as R^2 approaches 1. It is clear from this analysis that the demographic variable to choose is one where the association between it and the accident statistic of interest is the highest.

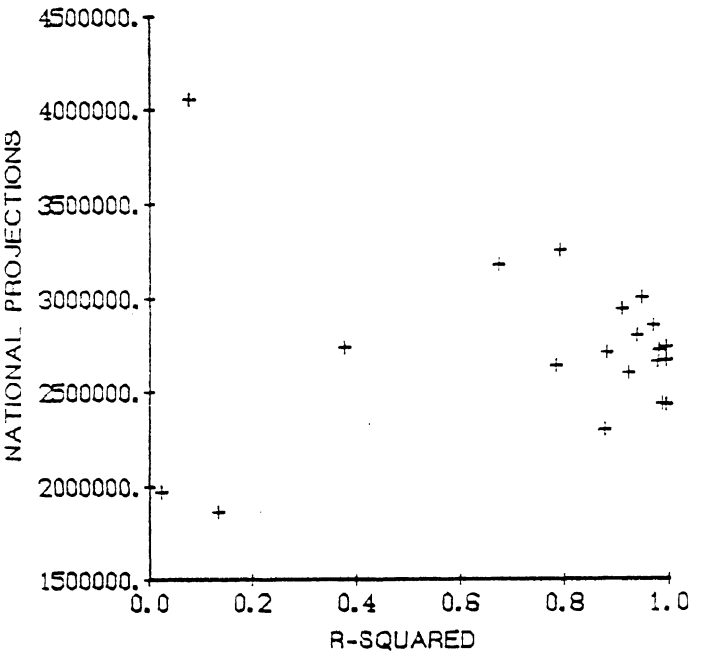

FIGURE 4.13 A Plot of The National Projections by the R^2 for the Regressions Used to Develop The National Projections

4.6.3 Modification for Missing Data. All of the examples in the previous subsections use variables that have no missing data. If the proportion of missing data is small (perhaps less than 5%), some analysts would ignore missing data and proceed with the analysis under the assumption that the small percentage of missing data would not alter their analysis significantly. In the NCSS data many of the variables have much more than 5% of the cases for which the information is missing.

An assumption underlying the national projection technique was the availability of unbiased estimates for the population totals for the units observed. The technique can be modified to adjust for missing data, in the case where data were not available for certain units or were lost in the sampling procedure. The adjustment involves obtaining new estimates for county totals that are adjusted for missing data. The method of adjustment of county totals is usually based on subjective knowledge or from a re-sampling of cases with missing data in a further attempt to obtain the data. Missing data procedures in general are

discussed in Section 4.1.4. In Section 4.5 the character of the missing data of key NCSS variables is examined.

The adjusted county total is then used to develop the predictive models for the unobserved units. This estimate of the population total within a county, if an imputation method were used to adjust for missing data, can be expressed for each county as

$$(4\text{-}35) \qquad t_i^* = t_i + u_i$$

where

t_i^* is the adjusted estimate of the i^{th} county total,
t_i is the observed total in the i^{th} county, and
u_i is the imputed total for the data missing in the i^{th} county.

If a reweighting procedure is used to adjust for missing data, the adjusted estimated county total can be expressed for each county as

$$(4\text{-}36) \qquad t_i^* = u_i t_i$$

where

t_i^* is the adjusted estimate of the i^{th} county total,
t_i is the observed total in the i^{th} county, and
u_i is the reweighted adjustment factor for the i^{th} county.

These adjusted t_i^* are used in place of t_i and the best fitting linear regression model with one independent variable is obtained by choosing $\hat{a}$ and $\hat{b}$ to minimize

$$(4\text{-}37) \qquad \sum_{i \epsilon s} (t_i^* - a - bx_i)^2$$

as in Equation 4-27.

If an imputation method is used and t_i^* is given by Equation 4-35 the estimate of the national projection is given by

$$(4\text{-}38) \qquad \theta(s) = \sum_{i \epsilon s} (t_i + w_i) + \sum_{i \epsilon s} [\hat{a}_1 + \hat{b}_1 (X_i - \bar{x}_s)]$$

where $\hat{a}_1$ and $\hat{b}_1$ are given by

$$(4\text{-}39) \qquad \hat{a}_1 = \sum_{i \epsilon s} t_i + w_i)/n$$

$$(4\text{-}40) \qquad \hat{b}_1 = \frac{\sum_{i \in s} (t_i w_i)(X_i - \bar{x}_s)}{\sum_{i \in s} (X_i - \bar{x}_s)^2}$$

If a reweighting procedure is used then the estimate of the national projection is given by

$$(4\text{-}41) \qquad \hat{\theta}(s) = \sum_{i \in s} t_i w_i + \sum_{i \in s} [\hat{a}_2 + \hat{b}_2 (X_i - \bar{x}_s)]$$

where $\hat{a}_2$ and $\hat{b}_2$ are given by

$$(4\text{-}42) \qquad \hat{a}_2 = \sum_{i \in s} t_i w_i / n$$

$$(4\text{-}43) \qquad \hat{b}_2 = \frac{\sum_{i \in s} (t_i w_i)(X_i - \bar{x}_s)}{\sum_{i \in s} (X_i - \bar{x}_s)^2}$$

Due to the nature of imputation methods at the present time there is no method that can be used to calculate the variance of the estimate. Rubin[46] advocates the use of multiple imputations, or multiple weight adjustments, to calculate the variability in an estimate due to the imputation procedure. This method will assess the sensitivity of the national projection to the imputation procedure but does not incorporate the variability of the county totals due to sampling.

4.6.4 Alternative Methods. In this subsection the national projections presented in Section 4.6.1 are compared to a method commonly in use to generate nationally representative numbers from the NCSS data. This method involves applying an inflation factor to a NCSS aggregate total. The inflation factor is the ratio of a population statistic to the comparable NCSS aggregate statistic.

---

[46] D. B. Rubin, "Multiple Imputations in Sample Surveys--A Phenomenological Bayesian Approach to Nonresponse," Imputation and Editing of Faulty or Missing Survey Data (Washington, D.C.: Department of Commerce, Bureau of the Census, 1978).

Kahane[47] and O'Day and Kaplan[48] have both used variations of this method. Kahane used, to form the inflation factor, the ratio of U.S. fatalities to the total number of fatalities in NCSS. O'Day and Kaplan based the inflation factor on the ratio of 1970 U.S. population to the population represented by the NCSS areas obtained from the County and City Data Book[49]. In Table 4.41 the actual inflation factors are presented. There are three inflation factors based on fatalities presented in this table. In Kahane's paper a preliminary version of the NCSS file was used. This data is referred to as Version 1. A near complete version of the NCSS data, that had approximately 55 accidents missing and that was used to produce the national projections in this report, is referred to as Version 2. The third inflation factor in the table is based on the number of fatalities in the officially released version of the 15 month NCSS file. The total number of fatalites for the 15 months of NCSS was obtained from the 1977 and 1978 FARS files. There were a total of 30,562 fatalities in towed passenger cars for the 15 month period of NCSS. There were 25,471 fatalities in 1977 and 5,091 fatalities in the first three months of 1978.

A national projection for the total number of accidents involving a towed passenger car is presented in Table 4.33. The inflation factors were applied to the estimate for the NCSS aggregate of the number of accidents involving a towed passenger car (31,867[50]). These inflated numbers are presented in Table 4.41.

---

[47]C. J. Kahane, An Evaluation of Standard 214. Technical Report (Washington, D.C.: National Highway Traffic Safety Administration, September 1979), Report No. DOT-HS-804-858.

[48]J. O'Day and R. Kaplan, The FARS Data and Side-Impact Collisions (Warrendale, Pa.: Society of Automotive Engineers, 1979), Report No. SAE 790736.

[49]U. S. Bureau of the Census, County and City Data Book, 1977: A Statistical Abstract Supplement (Washington, D.C.: Government Printing Office, 1978).

[50]Leda Ricci, ed., NCSS Statistics: January 1977-March 1978, Report No. UM-HSRI-79-80. Sponsored by the National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C., under Contract No. DOT-HS-8-01944 (Ann Arbor:Highway Safety Research Institute, The University of Michigan, October 1979).

The estimate produced using the inflation factor based on population gives the most conservative number of accidents. The estimates based on NCSS fatalities (Versions 2 and 3) are approximately the same. But these estimates are about 20% less than the estimate using the projection method. The very early inflation factor (Version 1) predicts a higher number of accidents than the national projection. The inflation factor method uniformly inflates all areas at the same rate. Rural and urban differences are ignored. The aggregate of all the NCSS areas is assumed to be perfectly representative of the entire U.S. The national project method uses a different inflation factor for each county so that the mix of urban and rural counties has less of an effect on the national projection.

There is no way to tell which of the methods is closer to the true population totals. When there is enough data (excluding rare events) the method of national projections seems to intuitively give better estimates. These estimates are more complicated to calculate and possibly the inflation factor method is useful in giving quick numbers that may be reasonable. Care must be taken when using the inflation method to use the most accurate data available to form the inflation factor as the estimate can vary quite substantially as seen in Table 4.41.

4.6.5 Summary. An estimation procedure, called a national projection, has been proposed for use in producing nationally representative statistics from the NCSS data. In spite of the fact that this estimate is biased, it has an intuitive appeal. The national projection combines the estimated total for the observed part of the population with a prediction about the magnitude of total for the unobserved part of the population. The prediction makes use of auxiliary information available for the entire population. It involves developing a model to predict from the auxiliary information the unobserved total of the statistic of interest. This model is developed using the data of the observed portion of the population and then it is used to predict the value of the statistic for the unobserved population.

TABLE 4.41
Comparison of Different Methods to
Adjust NCSS to Obtain
Nationally Representative Statistics
January 1977 to March 1978

Method	Number of Fatalities Represented	Ratio	ACCIDENTS
NCSS Fatal Occupants Version 1	300	101.87	3,246,291
NCSS Fatal Occupants Version 2	485	63.01	2,079,397
NCSS Fatal Occupants Version 3	500	61.12	1,947,711
1970 Population		51.09	1,628,085
National Projection			2,447,082

If the relationship in the selected areas between the accident statistic and the auxiliary information is the same as the relationship in the unobserved areas is a reasonable assumption, then there are two remaining main sources of bias. One term represents the degree to which the data fit the theoretical population model. The second involves the slope of the linear model, the number of unobserved counties, and the difference between the population mean and sample mean of the auxiliary variable used in the prediction model. With the NCSS design it appears that the second source of bias may cause an underestimate of the population total. Investigation of the choice of counties indicates that exclusion of any one of the selected counties causes the national projection to vary by no more than 4.5%.

Underlying the production of a national projection is the development of a model, based on the observed data, to predict the accident statistic. The sensitivity of the national projection on the choice of the auxiliary variable to be used for the projection was investigated. The national projections calculated separately for twenty auxiliary variables varied by about 17%. For variables with high R^2 there seemed to be less variability but no consistent trend. Excluding regressions with R^2 less than .6 the national projections varied by about 9.4%. Extreme predictions occurred when R^2 was low.

The national projections are compared with an alternative commonly used method that uses an inflation factor. These methods differ markedly. The use of an inflation factor produces an estimate consistently lower than the national projection. An exception is the early version of the NCSS file where 40% of the fatalities were not in the data file. The inflation factor method based on fatalities produces estimates approximately 15% less than the national projection.

Based on the analyses presented in this subsection the inflation factor method using fatalities will provide a quick method to produce nationally representative numbers. These may underestimate the national level somewhat. If what is needed is a good assessment of a national accident statistic the national projection method, although more difficult to develop, is appropriate. In summary, modifying the NCSS aggregate data using the national projection method takes more effort and as a consequence will yield a good estimate of the national experience. The inflation factor method will provide a quick method to get nationally representative statistics but appears to be sensitive to the ratio used in the inflation.

4.7 Significant Results. The NCSS design specified a purposive sample of seven areas. Within each area a stratified cluster sample of accidents was taken, except in two areas where a simple random sample of accidents was taken. For vehicles and occupants all within-area sample designs are clustered by accidents. Estimates for the aggregate of the NCSS areas were calculated. These NCSS accident statistics are organized and published in three publications:

1. NCSS Statistics November 1979.
2. NCSS Statistics: Passenger Cars June 1980.
3. NCSS Statistics: Light Trucks and Vans June 1980.

There are some key points to be made about these publications. The statistics in these publications describe police-reported accidents involving towed vehicles for the aggregate of the seven areas. There is reason to believe that the accident experience of light trucks and vans in the aggregate of the seven areas is not described nearly as well as the accident experience of passenger cars. So, in general, the tow-away accident population for light trucks and vans is not directly comparable with that of passenger cars. Finally, it should be noted that missing data counts have not been excluded from the tabulations and, consequently, the percentages for a particular category may be slightly underestimated.

Investigation of the NCSS data indicated that some accidents were sampled erroneously and some were excluded from the data file. During the sampling process, accidents were included that were not legitimately chosen accidents. The effect of these errors is probably small. This may indicate that accidents were also excluded in the process of sampling accidents but the extent of this is unknown. The excluded accidents were those with fatalities. Fatalities were studied separately and a comparison of NCSS with FARS and various state census files of accidents suggested that possibly as much as 20% of the fatalities occurring in the NCSS sites were missing from the NCSS data file. This would have the effect of making NCSS derived statistics underestimates of the population proportions.

As part of this analysis, variance estimates, or sampling errors, were calculated for selected statistics based on the Phase 1 data. This analysis was done in order to assess the magnitude of the effect of the cluster design on the variance associated with various NCSS statistics. The problem of presentation of sampling errors for publications like NCSS Statistics is given some attention in this analysis.

It was found that the sampling errors associated with some accident statistics were significantly affected by the cluster design. The confidence intervals based on the appropriate sampling error can be two or three times larger than a confidence interval based on sampling errors calculated for a simple random sample. The sampling errors associated with some injury level variables were actually less than the sampling errors under the assumption of simple random sampling due to gains from the stratification used in the NCSS design. It does not appear that using variances calculated on the assumption of simple random sampling will be a good approximation for all the NCSS statistics, and that to get more realistic estimates more complicated calculations are necessary.

Sampling errors are presented in two ways. A graphical representation of sampling errors was used. This method looks at the estimate as a function of its sampling error. This method easily summarizes a great deal of information. Design effects can also be used to categorize certain types of statistics, like accident, vehicle, or occupant level statistics or even a finer breakdown, that summarizes the effect of the design on the sampling error. So if the design effect is available it can be multiplied by the simple random sample variance and the result used as an approximation to the appropriate sampling error. Both procedures are used in subsection 4.4 to present the sampling errors calculated.

The focus of the analysis of missing data was on differences between the vehicles or occupants that were missing data and those that had complete information. This is the first step in finding appropriate missing data adjustment procedures. The literature describes various procedures that are currently being used. They all assume that there is no basic difference between missing and non-missing data elements. That

is, only the proportions within the categories of a variable differ between the two segments. To develop missing data adjustment procedures some assessment must be made of how these two segments differ.

The key variables involved in the analyses were OAIS and Delta V. OAIS is an ordinal variable that indicates a degree of injury severity, and Delta V is a measure of crash severity. Missing data rates were calculated for OAIS and Delta V and the rate at which one or the other was missing. The missing data rate for OAIS is about 30%. Calculation of a missing data rate for Delta V should be relative to those vehicles where Delta V could be calculated. Using this base, the missing data rate for Delta V is about 34%. The rate at which one or the other is missing is about 70%.

Missing data for fatal and non-fatal occupants were investigated separately. Information was obtained from the hard copy documentation for 333 fatal occupants and 388 non-fatal occupants. For the fatal occupants with missing data there was a higher proportion of occupants with OAIS coded with a maximum injury. There was also an increase in the proportion of vehicles with fatal occupants at higher categories of Delta V. For non-fatal occupants there was a substantial difference in the proportion of minor injuries.

A procedure was developed to produce nationally representative estimates. This procedure uses NCSS statistics and demographic variables by county. A model relates the NCSS statistics and the demographic variables. This model is then used to predict the statistics for the counties not observed. These predictions are then combined with the NCSS statistics to form national projections.

The procedure to develop national projections can be used for univariate and bivariate distributions as well as simple statistics. The procedure appears not to be very sensitive to the exclusion of a particular county from the modelling stage of the analysis, the national projections vary by about 5%. Investigation of the sensitivity of the national projection to the choice of demographic variables indicated that when demographic variables were chosen where the correlation was significant (in this analysis greater than .75) the national projection varied by about 9%. This procedure works well as long as the

correlation between the accident statistics and demographic variable is high. This method does not produce stable projections when looking at events with a low probability of occurrence or with variables that have large amounts of missing data.

All of the NCSS statistics for which national projections were developed involved only variables with very little missing data. The procedure is modified to allow for different types of missing data adjustments. The modification involves using the adjusted county statistics in the analysis rather than the unadjusted totals directly from NCSS. This will inflate the national projections as would be expected.

Finally, the national projections based on the method developed in this project are compared with a frequently used procedure. This procedure involves inflating the NCSS aggregate by an inflation factor. One factor of this type used is the ratio of U.S. fatalities to NCSS fatalities. The inflation estimates were lower than the national projection of the total number of accidents involving a towed passenger car. When an approximate estimate is needed the inflation method will provide easily obtainable estimates. But, when good national estimates are needed the national projection method is appropriate.

5 ACCIDENT ANALYSIS MODELS

The NCSS data contain a wealth of information on the current highway accident experience. Ultimately, one would also hope that analysis of this information would reveal modifications which would reduce future highway accident losses. Accident analysis models are one attempt to synthesize current information in a manner which is intended to aid highway safety decisions. These models use information on the current accident experience to estimate the potential for injury reduction of proposed restraint systems.

The objective of this chapter is to review the implications of the analytical work described in this report for the use of the NCSS data in current accident analysis models. The focus of this effort is the Kinetic Research Accident Environment Simulation and Projection Model (KRAESP)[51]. NHTSA is currently funding Kinetic Research to incorporate the NCSS data in the KRAESP model as part of "Basic Ordering Agreement for Systems Engineering Studies," Contract No. DOT-HS-9-02096. The first subsection describes accident analysis models in general, and the KRAESP model in particular, focusing on the role of the accident information which is used. This subsection is followed with a discussion of our analysis of the NCSS data and a discussion of the implications of this work for the use of the NCSS data in the KRAESP model.

5.1 Existing Models

This subsection begins with a brief discussion of the general objectives of accident analysis models, and is followed by a longer description of the KRAESP model focusing on the role of the accident data which is required.

5.1.1 General Objectives. Currently there are more than 150 models of the motor vehicle transportation system in the public

---

[51] D. Redmond and K. Friedman, "Introduction to the Kinetic Research Accident Environment Simulation and Projection Model," Prepared under DOT Contract No. DOT-HS-9-02096, Kinetic Research Draft Report No. KRI-TR-041, January 1980.

domain[52]. Virtually all of these focus on economic forecasting of demand in terms of vehicle miles traveled. An example of these models is the Automobile Demand Model developed by Wharton Econometric Forecasting Associates[53]. However, a few models have been developed which seek to forecast the motor vehicle accident experience. An early model of this type developed by Joksch[54] used time-series methods to predict the number of fatalities as a function of the vehicle size mix.

Both of the models of interest to this discussion were developed as part of the Experimental Safety Vehicle programs sponsored by the Department of Transportation during the early 1970's. Ford Motor Company developed the Safety System Optimization Model[55], and Minicars, Inc. developed the Research Safety Vehicle Accident Analysis Model[56], which later evolved into the Kinetic Research Accident Environment Simulation and Projection (KRAESP) model. The objective of these models is to estimate the reduction of deaths and injuries for future populations of vehicles with various proposed occupant protection systems.

In general, these accident analysis models must synthesize the current information in both the accident causation and the vehicle

---

[52]Richardson, B.C., Segel, L.D., Barnett, W.S., and Joscelyn, K.B., An Inventory of Selected Mathematical Models Relating to the Motor Vehicle Transportation System and Associated Literature. HSRI, 1979, Report No. UM-HSRI-79-37. Ann Arbor, Michigan: UMI Research Press, an imprint of University Microfilms International. Sponsored by the Motor Vehicle Manufacturers Association.

[53]G.R. Shrink and C.J. Loxely, An Analysis of the Automobile Market: Modeling the Long-Run Determinants of the Demand for Automobiles, Final Report to the Department of Transportation , Transportation Systems Center, February 1977.

[54]H.C. Joksch, An Accident Trend Model - Final Report. Center for the Environment and Man, Inc. Hartford, Connecticut, March 1975.

[55]Research Safety Vehicle (RSV) Phase I Final Report, Volume II, Ford Motor Company. Prepared for the Department of Transportation under Contract No. DOT-HS-4-00842, June 1975. DOT HS-801 599.

[56]D. Struble and G. Bradley, Research Safety Vehicle Phase I, Volume II, Program Definition Foundation, Minicars, Inc. Prepared under Department of Transportation Contract No. DOT-HS-4-00844, June 1975. DOT HS-801 604.

crashworthiness areas. The two major functions of current accident analysis models are:

1. Project the accident experience (causation).

2. Project the injury response (crashworthiness).

Projecting the accident experience requires that information on the size, composition, and use of the current and hypothetical vehicle and occupant population be supplied as input. Also, a statistical description of the accident experience of the current population is required. Current knowledge on the relationship of the characteristics of the vehicle and driver populations to their resulting accident experience (causation) is then used to estimate the accident experience of the hypothetical population of vehicles and drivers. The result of this computation is a complete description of the number and kind of accidents the hypothetical population of vehicles and occupants will be involved in.

Projection of the injury experience requires an analogous treatment. Having estimated the accident experience of the hypothetical vehicle and occupant population, information on the probability of injury and death for both the current and the proposed occupant protection systems is used to estimate the injury experience of the hypothetical population. Again, the necessary input for this computation is the injury experience of the current occupant population. Current knowledge of the relationship of the collision conditions to the resulting injuries and deaths is used to relate the performance characteristics of the hypothetical occupant protection systems to their expected injury experience.

5.1.2 The KRAESP Model[57,58,59,60]. This subsection is intended to provide a brief overview of the Kinetic Research Accident Environment Simulation and Projection (KRAESP) model. The focus of this overview is the role of the accident data inputs to the model. Consequently, other major operations of the model may be omitted or given only passing mention.

In the KRAESP model, the accident experience is divided into numerous subsets. Injury estimation is carried for each subset, and the results summed. The model, then, is defined by the variables and levels which define the subsets. These are listed below:

Accidents are subset by:

calendar year

Vehicles are subset by:

manufacturer
model year
weight class
restraint system by occ. seat location

Occupants are subset by:

collision severity
collision mode (vehicle-to-vehicle, single vehicle)
damage area (by clock direction)
occupant seat location
(occupant age)
(body region of injury)

---

[57] D. Redmond and K. Friedman, "Introduction to the Kinetic Research Accident Environment Simulation and Projection Model (KRAESP)," Prepared under Department of Transportation Contract No. DOT-HS-9-02096. Kinetic Research Draft Report No. KRI-TR-041, January 1980.

[58] K. Friedman, R. Thomson, and D. Redmond, "The Kinetic Research Accident Environment Simulation and Projection Model," prepared under Department of Transportation Contract No. DOT-HS-7-01552. Minicars, Inc. Draft Report No. KRI-TR-027, July 1973.

[59] D. Struble and G. Bradley, Research Safety Vehicle Phase I, Volume II, Program Definition Foundation, prepared under Department of Transportation Contract No. DOT-HS-4-00844, June 1975 DOT HS-801 604.

[60] N. DiNapoli, et. al., Research Safety Vehicle Phase II, Volume II, Comprehensive Technical Results, Minicars, Inc., prepared under Department of Transportation Contract No. DOT-HS-5-01215, November 1977.

Occupant age and body region of injury are subset variables which are available in the current version of the KRAESP model, but which have not yet been employed for lack of input data.

Figure 5.1 shows the overall organization of the KRAESP model and the locations of the accident data inputs. The KRAESP model requires inputs over a period of calendar years, and produces projections for each year. This discussion is limited to the data requirements for a single year for simplicity. The first box indicates the vehicle population projections. The necessary input is the number of vehicles introduced during the model year by manufacturer and weight class. This information is combined with the information on the vehicle population at the beginning of the year and estimated scrappage rates to project the total population for the current year. The restraint systems may be specified by seat location and impact mode (front, side, rear, and roll). In other words, the model allows different restraint systems to be specified for the different impact directions; for example, an airbag for frontal impacts, headrest for rear impacts, and, perhaps, none for side impacts.

Accident data are input in the second box which describes the occupancy of each seat location and the closing speed distribution for each impact mode (vehicle-to-vehicle and single vehicle) and damage area (clock direction). The information in boxes one and two is combined to produce the description of the projected "accident environment" represented by the third box. A key assumption is involved at this point in the estimation of the number and distribution of accidents. The total number of accidents is adjusted in proportion to the estimated number of vehicle-miles. Vehicle mileage, in turn, is assumed to be only influenced by scrappage (vehicles taken out of use) and vehicle age. None of the accident information supplied in box two is altered. Only the total number of accidents is adjusted. One way of viewing this projection procedure is that it effectively controls (eliminates) any possible changes in the probability of an accident as a result of vehicle size or use. Consequently, any changes predicted are the result of the crashworthiness analysis, and year-to-year changes simply reflect

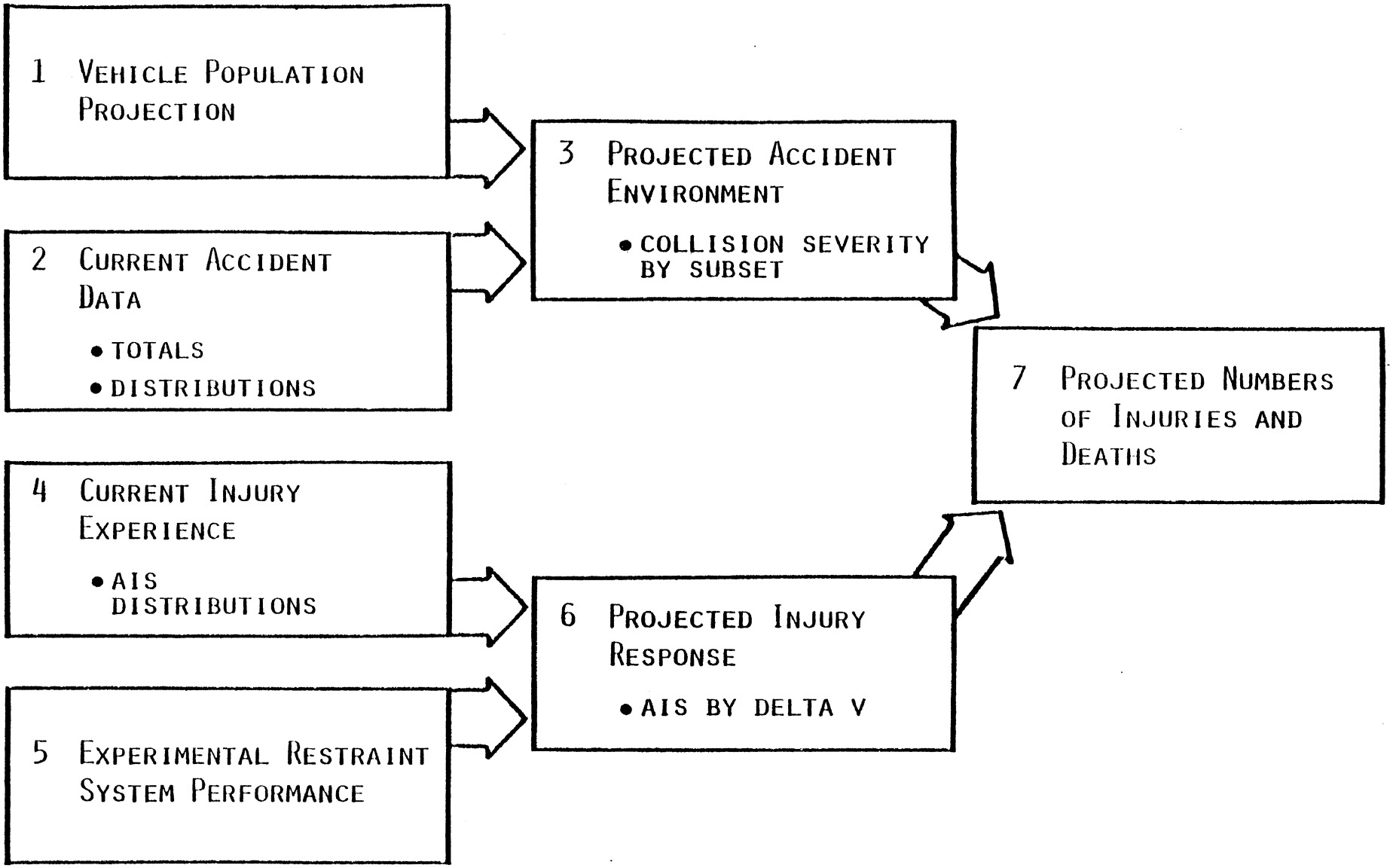

FIGURE 5.1 SIMPLIFIED DIAGRAM OF THE KRAESP ACCIDENT ANALYSIS MODEL

the growth of the vehicle class of interest in relation to the total population of vehicles.

A key analysis variable in this portion of the model is vehicle weight, which is generally categorized into "classes." Vehicle weight is central because of its interaction with collision severity. Recall that the accident data provided in box two included distributions of closing speed for each impact mode and damage area. Closing speed is the velocity of the struck vehicle with respect to the striking vehicle or object. These distributions are felt to be relatively independent of the vehicle population (although they may be influenced by oil shortages, speed limits, or other major changes in the automobile transportation system). The measure of collision severity required for the crashworthiness analysis (boxes 4-6) is the velocity change of the vehicle, Delta V. Basic momentum principles indicate that the velocity change of each vehicle (in a two-vehicle collision) is related to the closing speed and the masses of the vehicles by the following equation.

$$\text{Delta } V_1 = V_c(M_2/(M_1 + M_2))$$

where:

V_c is the closing speed, and
M_1 and M_2 are the respective vehicle masses.

Having specified the weights of the vehicle population in the first box, the above relationship is used to compute the distributions of Delta V for each weight class given the distributions of closing speed for each impact mode and damaged area provided as part of the accident data in box two.

The model must now estimate the injury response of the proposed restraint systems for the specific accident subsets used in steps one through three. The input data for the proposed restraint systems is in the form of engineering tests. The major hurdle in this portion of the model is the relationship of the engineering test measurements to the probability of injury or death. To accomplish this task, accident data describing the current injury experience (for unrestrained occupants) is provided to the model as shown in the fourth box. The unrestrained case is also included in the engineering tests. The accident data supplied

takes the form of distributions of AIS (Abbreviated Injury Scale) levels for each impact mode, damage area, occupant seat location, and 5 mph increment of Delta V. For each increment of Delta V for the proposed restraint system, the engineering test measurements (deceleration, for example) are used to identify a comparable severity level for the unrestrained case. The distribution of AIS from the accident data for this Delta V level is then identified as the injury response of the proposed restraint system. The results of this process are the projected injury response functions shown in the sixth box in Figure 5.1. Separate response functions are computed for each impact mode, damage area, and occupant seat location.

Having projected the numbers of occupants exposed to each of the accident subsets in the projected accident environment (box 3) and the injury response of the proposed restraint systems under each of these same conditions (subsets), this information can be readily combined to estimate the numbers of injuries and deaths as shown in the final box in Figure 5.1.

In view of the development of mechanistic models presented in Section 3 of this report, the major question to be addressed at this point in the model is whether collision severity (Delta V) is a sufficient predictor of the probability of injury or death (more specifically, the probability of a given AIS level). This material is presented in the following subsection.

5.2 The NCSS Data

This subsection identifies the work presented in this report which is relevant to the application of the NCSS data to the KRAESP model. Our efforts have been presented under two broad headings, population statistics and mechanistic models. Both of these topics are relevant to the application of the NCSS data to the KRAESP model.

Earlier in this section, the two major tasks of the accident analysis models were identified as the projection of the accident environment and the projection of the injury response. Each of these tasks, in turn, required accident data inputs. In the projection of the accident environment, the distribution of accidents across subsets is

required. In the projection of injury response, the distribution of AIS is required for each subset. For each of these, national estimates of the accident distributions are required.

The aggregate of the NCSS areas does not provide national estimates since the NCSS data collection sites were selected purposively. The development of national projections from the NCSS statistics is presented in Section 4.6 of this report. The method uses county-level demographic information which is published for all counties in the United States. Models are then developed relating NCSS accident statistics to the county-level variables. These relationships are used to project the accident statistics to the counties not included in the NCSS program. The variance of this estimate is also examined.

Application of this method to the data requirements of the KRAESP model would be difficult. In general, the technique is not suitable for all accident statistics. The necessary models must be developed separately for each statistic. Some statistics, especially those describing small subsets of the data, may not exhibit strong relationships with any of the available demographic data. In these situations, the resulting national projection is more likely to be biased. Also discussed in Section 4.6 is the simple ratio, or inflation, projection method. It appears that this method may also produce somewhat biased estimates.

However, a careful application of these methods in conjunction with suitable supplementary information may yield usable results. For example, population models might be used to estimate totals for major subsets, and then ratio estimates could be applied for the smaller cells. For fatal accidents, the FARS file might provide good totals for ratio estimates. In summary, the NCSS cannot by itself easily provide all the required national estimates, but NCSS can serve as a valuable data source.

The mechanistic models presented in Section 3 of this report are also relevant to the KRAESP model. In the projection of injury responses for the proposed restraint systems, distributions of AIS derived from accident data for unrestrained occupants are adjusted to estimate the distributions of AIS for the proposed restraint system

based on a comparison of the results of engineering tests and the collision severity measure, Delta V. The underlying assumption here is that the resulting AIS distributions can be adequately predicted by collision severity (Delta V). This assumption will be met to the degree that the mathematical model used accurately reflects the true physical relationship between the relevant variables. Models which attempt to reflect the governing physical principles have been labelled "mechanistic" models in this report. The following paragraphs will compare the results of our attempts to develop mechanistic models with the KRAESP model.

Before beginning this discussion, it should be pointed out that the problem being addressed here is basically that of vehicle crashworthiness. Given that a collision has occurred, the objective is to relate variables which describe the nature of the collision to the injuries which result. If this can be done successfully, then it is reasonable to expect to predict the change in the resulting injuries when the characteristics of one of the elements in the model, like the restraint system, are changed. Such an expectation is reasonable when the model accurately reflects the physical mechanisms which govern the collision event. A similar problem is presented in trying to estimate the number and type of accidents to be experienced by a hypothetical population of vehicles. Here, the more complex problem of accident causation is involved. In projecting the accident environment, the number of accidents is adjusted in proportion to the estimated number of vehicle miles. Ideally, these adjustments would also be based on physical principles relating the type and use of vehicles to the resulting accident experience.

If vehicle populations are hypothesized with appreciably different distributions of car size, for example, then it might also be reasonable to envision that the use of these vehicles might be different. Smaller cars tend to be used more in urban environments. The accident experience of vehicles in urban usage is appreciably different from those in rural usage. Collision severity is lower for urban accidents, as is the proportion of single-vehicle accidents. Other trends unrelated to the vehicle population might also be important. These

might arise due factors such as oil shortages or changes in national speed limits.

Adjusting the distributions in the national accident experience to reflect year-to-year changes or trends is a subset of the larger problem of accident causation. Although the accident analysis models were originally developed to address the crashworthiness area, the KRAESP model has been extended to make projections in the accident causation area. The user has the option to provide inputs on the type of brake system for each vehicle. Each system is, in turn, presumed to modify the distribution of closing speeds in the accident population. While this seems a plausible first approximation, it opens up a vast and relatively unknown area.

Projections in the causation area will only be as good as the models used to make the projections. The validity of the accident causation models will rest on their foundation in the physical principles which determine the occurrence or non-occurrence of an accident. This treatment of the accident causation process is felt to be far more difficult then the crashworthiness area. Unfortunately, assumptions in this area are currently required in the accident analysis models just to adjust the national accident experience to reflect changes in the total estimated vehicle use resulting from changes in the vehicle population. The total number of accidents is adjusted to reflect estimated changes in total vehicle mileage. No differences in vehicle mileage are included for vehicles of differing sizes. Only the total number of accidents is adjusted, so that the distributions remain unchanged. These assumptions imply an accident causation model. The remainder of this discussion deals with crashworthiness.

The basic model used by KRAESP to project the injury response of the proposed restraint systems was summarized in Section 5.1.2. The accident experience is first subset, and then particular distributions of AIS are associated with 5 mph increments in Delta V. The subsetting for development of mechanistic models presented in Section 3 followed similar lines. The initial subsets separated front and side impacted vehicles. Back damaged vehicles were not studied, and an examination of rollover vehicles revealed only that ejection was closely associated

with severe injury. The front and side groups were further divided into vehicle impacts and fixed object impacts. For the front groups, each of these were further separated into impacts to the center-front and impacts involving the right- or left-front portion only. In the side impacts, impacts to the passenger compartment were separated from non-passenger compartment impacts. Near-side and far-side occupants were treated separately in the side impacts, as were drivers and right-front occupants in the front subsets. Other seat locations were not included in the front subsets.

In the KRAESP model, damage area is identified by a "clock position" which "points" to the location of the contact with the other vehicle or object. The "clock direction" coded as part of the Collision Deformation Classification[61] in the NCSS data is defined differently. In the NCSS data, and in the CRASH program, this clock position is used to identify the principal direction of force. As such, it identifies the vector direction of the relative velocity with the struck object, and the direction of the resulting Delta V. The location of damage is identified by subsequent characters in the CDC. This situation poses no conceptual problems in applying the NCSS data to the KRAESP model; it simply explains why our identification of damage location is specified somewhat differently.

With these subsets, Delta V was indeed the strongest predictor of injury severity. However, the examination of residuals presented in Section 3 indicates that, although the overall percent of correct predictions was reasonably good (80-90%), the prediction of severe injuries was frequently less than 50%. Addition of variables such as occupant age, intrusion, and accident location (rural/urban) helped somewhat, although the magnitude of their effects was appreciably less than that of Delta V.

Further examination of the mispredictions revealed that addition of body region, in particular, and injury type secondarily, substantially reduced the mispredictions. Even for the severe injuries, the percent

---

[61]"Collision Deformation Classification--SAE Recommended Practice J224a," SAE Handbook, 1980 Ed. (Warrendale, Pa.: Society of Automotive Engineers, 1980), pp. 34.109-34.113.

correct prediction was increased to 70% or more. This finding poses some problems, however. Putting injury type into the model is a little like putting the dependent variable in as an independent variable, since many injury types can only be assigned to one or two AIS levels. Similarly, there is a strong correlation between body type and injury type (extremities usually incur fractures, concussion can only occur to the head, etc.). While it is informative to know the source of the mispredictions, it is not clear what to do. Ideally, one would like to find variables which would predict the body region. It would seem that contact points and/or direction of principal force might be useful. Efforts to use this information were not particularly successful. Alternatively, separate models might be developed for each generalized body region. This problem was recognized by Klimko,[62] who recommended that at least a maximum Occupant Injury Classification (OIC) and contact point be coded for each of at least three generalized body regions including an indication of no injury and/or no contact. We would agree with this recommendation, although the data collection problems would be difficult indeed. Current efforts to get contact points are only marginally successful. Our modelling efforts seem to point to the need to treat body regions separately. Currently, the KRAESP model has the capability to subset by body region, although this capability has not been used. Subsetting by body region would seem to be the next logical step.

Another critical assumption in the projection of injury response in the KRAESP model is the use of engineering test measurements, such as chest deceleration, to establish equivalence of injury severity in the accident data and the response of proposed restraint systems. While this topic is also outside the scope of this discussion, it would seem that subsetting on body region would somewhat improve this situation, since it would then be possible to ensure that the engineering measurement would be taken from the same generalized body region as that of the observed injuries in the accident data.

---

[62] L. Klimko and K. Friedman, Statistical Analysis of Crash Conditions and Their Relationship to Injuries, Kinetic Research, Inc., Final Report June 1978.

From a practical standpoint, missing data is one of the most serious problems with the NCSS data. This topic has been discussed extensively already. Delta V and AIS are both present on only 40% of the cases. Even without worrying about possible bias, cell sizes are likely to be too small for many of the small subsets defined by the KRAESP model. Missing data on AIS is considerably reduced on the NEWOAIS variables generated by the NCSA algorithm. At a minimum, this algorithm could be modified to identify OAIS 0-2,3, and 4+. The loss of detail in the final output would probably be worth the reduction in missing data. Beyond this, it would seem necessary to develop some of the missing data adjustment techniques presented in Section 4.1.4 and discussed in Section 7.5.

Missing data on Delta V will also be difficult to deal with. Closing speed is not available in the NCSS data. To compute this variable it will be necessary to create a "two-vehicle" file which matches the vehicle-level information for the two vehicles used in the Delta V computation. This matching process can be carried out for nearly all cases, although some difficulty is involved. For this project, a two-vehicle file was only prepared for the preliminary data. Closing speed was not computed.

A related problem is the influence of the towaway threshold on the vehicles selected for NCSS investigation. Klimko[63] observed that small cars in the RSEP data actually had a slightly lower mean Delta V than larger cars. This is certainly contrary to what one might expect if these vehicles were involved in accidents having the same distribution of closing speeds as larger vehicles. Based on Equation 5-1, lighter vehicles would be expected to have higher average Delta V values. Klimko suggested that a higher Delta V might be required before a larger vehicle was towed from the scene, than for a smaller vehicle to be towed. If this were the case, more small cars with low Delta V's and fewer large cars with low Delta V's would be towed, and subsequently eligible to be case vehicles. A review of Delta V by car size in the

---

[63] L. Klimko and K. Friedman, Statistical Analysis of Crash Conditions and Their Relationship to Injuries, Kinetic Research, Inc., Final Report June 1978.

NCSS file revealed that the lighter vehicles do appear to have slightly higher Delta V values, on than average, than larger vehicles. Again, in order to pursue this a two-vehicle file is required.

From the standpoint of the influence of the towaway threshold, it would seem desirable to include non-case vehicles as well as case vehicles in the two-vehicle file. In that way, the closing speed would be obtained even if only one vehicle were towed. Distributions of closing speed would then be less sensitive to any variation in the towaway threshold with car size.

5.3 Summary

Two major points arise from this brief review of accident analysis models in general. They are:

1. National estimates of the accident experience are required.

2. The projections will be valid only to the degree that they reflect the actual physical principles and mechanisms which govern the events being simulated.

National estimates are required since it is the national accident experience which is being projected. The important point here is that statistically based national estimates (which will eventually be available from the NASS) carry with them estimates of their variance. If this information were carried through the simulation process, one would be in a much better position to evaluate the variability of the resulting projections.

The second point embodies the essence of what we have described as "mechanistic" models. For many applications a statistical description of the current situation is completely adequate. The situation is much different, however, when one wishes to project the effect of proposed changes in the system. Statistical correlations present before the changes are introduced may be altered. Controlled experiments generally cannot be conducted in a social system. The alternative is to ground the statistical models in the physical principles and mechanisms which govern the event being simulated. This is the critical issue in the projection of the accident experience of the hypothetical vehicle population, and also in projecting the injury response of the proposed

restraint systems. Projection of the accident experience of the hypothetical vehicle population basically involves models or assumptions in the area of accident causation. However, the current version of the KRAESP model does not appear to take into account the possibility that vehicles of different size classes may not be used in the same driving environment, and, consequently, may have different accident experience. For example, small vehicles may be used mostly in urban environments where they are typically involved in fewer single-vehicle accidents and, in general, collisions of lower severity. The implications of the assumptions currently employed in the projection of the accident experience appear to need careful review.

In the crashworthiness area (the projection of injury response), the subsets used by the KRAESP model are generally comparable with those which evolved from our work. The important observation here, is that the prediction of severe injuries was correct only about 50% of the time unless body region was included in the model. The implication is that separate models should be developed for at least three or more generalized body regions. Since not all injuries are coded for the NCSS data, separate injury distributions for each body region may be somewhat underestimated. We concur with the recommendation previously attributed to Klimko; that at least the most severe injury should be recorded for each of at least three generalized body regions, including the occurrence of no injury to the body region.

A final observation is that missing data will be a serious problem. Either Delta V or OAIS are missing on 60% of the file. Our modelling efforts only addressed the most promising front and side impact subsets. Alternative techniques will be needed where Delta V is not a suitable measure of collision severity. Statistical techniques such as those discussed in Section 4.1.4 or 7.5 will have to be employed to address the missing data problem.

6 CLINICAL WORK

A major purpose of the NCSS program has been to develop a data set which would permit predictive modeling of the relationships between crash severity (and type) and occupant injury. Such models are discussed in Section 3. In addition to such statistical procedures, however, the NCSS case reports are relatively rich in detail and offer the possibility of clinical review. This section of the report discusses the goals of the clinical review process, the methods, and the principal findings. These clinical studies have been published in separate project reports as well as in the scientific literature. Readers are referred to the original reports for more detail; summaries will be presented here.

As a result of the coding conventions used in the NCSS program some information which was available in the field reports was not coded into the computerized files. Such information ranged from informal injury descriptions (not fully supported by medical documentation) to photographs of crash damage (which contain more complete detail than could be coded into the Collision Deformation Classification system. Further, injury details in the written reports were usually provided on an anatomical diagram, and could be interpreted more fully by persons with medical training. In the following clinical studies much use was made of these documents.

While the NCSS field reports have been found valuable, the reviewers have noted some shortcomings which limited their usefulness. It is hoped that comments regarding the quality and completeness of the data and the reporting methods will serve to guide the acquisition of data in future programs. Consequently, one part of this section of the report will discuss this matter.

The five separate reports produced under the clinical task during this project include (1) a bibliographic review, (2) a study of particular side impact cases, (3) an analysis of cervical injuries, (4) an analysis of ocular injuries, and (5) an analysis of lower extremity injuries.

6.1 The Bibliography

As a first step toward the clinical studies, the recent biomechanics and automotive injury literature was reviewed. The purpose was to determine the state of knowledge relative to injury type, body region, frequency, severity, and cause, and to define those clinical areas in which the NCSS data might best fill gaps in the current knowledge. The bibliographic report was divided into six sections according to body region, as follows:

1. The head and face
2. The neck and throat
3. The thorax
4. The abdomen
5. The vertebral column
6. The extremities

Although the titles of many articles in the medical literature are enticing, most such papers are concerned with treatment plans, associated medical problems, and/or case histories of a very specific type of injury. There is usually little crash information contained within any of these articles. A typical crash description might read: "The patient...was an occupant of a car that hit a tree at high speed." Needless to say, such "crash data" do not add much to our understanding.

Biomechanics research laboratory data on human tolerance are generally very specific in terms of the imposed impact conditions, the body region impacted, and the subject kinematics. However, the test conditions may not be totally representative of field conditions, and may only consider a portion of the overall sequence of events in a real crash. Biomechanics research on injury is further restricted by the use of surrogates of the living human (cadavers and animals) as models to study the mechanics of trauma. Often the number of subjects tested in a particular study is necessarily small. The lack of large data samples is counteracted somewhat by the well-defined test conditions and the degree of control of the mechanical variables during a test.

The literature review suggested that the study of accident details in the NCSS program might bridge the gap between past medical studies and laboratory experimentation. A measure of crash severity, admittedly less precise than laboratory instrumentation, is available in the form

of a computer reconstruction (Delta V calcualated by the CRASH-2 program).[64] Most of the NCSS injury data have been acquired from qualified medical sources. This information, combined with a careful implementation of a sampling plan, should provide a better bridge between biomechanics, injuries, and injury causation than has been previously available.

While the biomechanics of severe head injuries are relatively well understood, knowledge of the frequency of such injuries, and of the sources (contacts) is not well defined. The sampling procedures used in NCSS should permit such frequency estimates.

Anterior neck (throat) injuries are occasionally described in the literature, but their frequency is not well reported. Again a major contribution of NCSS may be to provide that knowledge.

With regard to thoracic injuries, the NCSS data should provide both overall frequency information and a data base to answer specific question regarding occupant age, direction and type of loading, and crash conditions. Such information should complement presently available biomechanics research.

Probably the abdominal region has a greater lack of human tolerance data than any other area. While more biomedical/biomechanical research is needed, the NCSS data may be expected to complement this with frequency and crash condition information.

The cervical region of the vertebral column is of particular importance. The presence of a large number of autopsy reports in the NCSS data should permit a more complete understanding of the number and type of such injuries, as well as the crash circumstances which lead to them.

Some human tolerance data are available on the lower extremity, but since injuries to this region are seldom fatal, they have not received much attention. The NCSS data may be expected to contribute both

---

[64] R. R. McHenry and J. P. Lynch, "CRASH2 Users Manual." DOT/HS 802-106, November 1976

overall frequency information and more detailed injury descriptions than have been available to date.

Within the limits of time and funds available for this project, the literature review led directly to studies of three specific body regions--the eye, the lower extremity, and the neck. In addition, selected side impact cases were studied so as to compare the actual accident experience with typical laboratory side-impact simulations.

6.2 Side Impact Studies

Approximately 90 selected side-impact cases from the NCSS were studied to determine similarities and differences between actual crashes and laboratory (sled) crash tests. Sled tests simulating side impact have generally been conducted at a 90^{0} impact angle, and the cases reviewed in detail were those with a near-side occupant and a reported 3 o'clock or 9 o'clock impact vector.

Of approximately 90 cases studied, 51 were judged comparable to the laboratory situation. The remainder generally involved cars struck at a point remote from the passenger compartment, often with considerable rotation of the vehicle. Injuries for the 51 cases were tabulated by crash severity (Delta V) and the conclusion was drawn that they were quite similar to those observed in laboratory (sled) tests at a slightly higher Delta V.

The report has been published separatedly as "Analysis of NCSS Side Impact Cases", by J. W. Melvin, D. H Robbins, D. F. Huelke, and J. O'Day, and was subsequently published as SAE Paper 800176. February 1980, presented at the SAE Congress and Exposition, 25-29 February, 1980.

6.3 Leg Injuries

Lower extremity injuries are identified in the NCSS data by body region (pelvis, thigh, knee, leg, and ankle/foot), by injury type (fracture, laceration, etc.), and by the source or point of contact

(instrument panel, foot controls, etc.). In the paper prepared on this topic, only injuries at AIS-3 or above were studied in detail.[65]

Analysis of the NCSS data indicates that injuries of the more severe nature (AIS 3 and 4) in the lower extremity are exceeded in frequency of occurrence only by those in the thoracic region. When national estimates are made, it appears that there are some 27,000 car crash survivors each year sustain the more severe lower-extremity injuries. This is approximately equal to the total number of passenger car occupants who are killed annually.

The medical consequences of lower-extremity injuries of the more severe nature may be extreme, including prolonged immobilization, long recovery periods, and the potential for the development of traumatic arthritis. Bone infection is a hazard that can cause bone weakening, recurrent infection, and life long threat of disability. Not infrequently many of the individuals with these AIS 3 and 4 lower-extremity injuries will have some degree of permanent impairment.

Front right passengers more often had the more severe lower extremity injuries than did other occupants; drivers sustained a lower than average frequency of the more severe lower limb injuries.

The more severe lower-extremity injuries are most often sustained by unrestrained occupants impacting objects in front of them, with the lower instrument panel being the main contact location. Fractures are the most common type of the more serious lower-extremity injuries.

The instrument panel is associated with injuries of the pelvis, thigh, knee, and leg, whereas the ankle/foot region is almost always injured by floor or foot control contacts. The back of the front seat and the side interior are the contact points most often listed for side impact crashes.

Of all of the sub-regions of the lower-extremity, the pelvic injuries are found most in drivers, whereas the front passengers had pelvis or thigh as the two areas most often injured.

---

65 "Lower Extremity Injuries in Automobile Crashes--an analysis of NCSS data,", by Donald F. Huelke, James O'Day, John D. States, and Thomas E. Lawson, report number UM-HSRI-80-10, January 1980.

Direct impact loading to any areas of the lower-extremity can cause injuries in that body region. In many cases force transmission through bone to other lower-extremity area can cause fractures and/or dislocations remote from the impact site. Compression or twisting forces, especially at the ankle area, are believed to be the main cause of the injuries to the ankle and foot.

Seat belt systems appear to reduce the more severe lower-extremity injuries; however, there are too few cases available in the NCSS data to make a definitive statement.

Increased attention to impact characteristics of the lower instrument panel may prove beneficial in reducing the occurrence of the more severe lower-extremity injuries.

6.4 Ocular Injuries

Tempered windshields commonly used in Europe have been shown to be highly related to ocular injuries. Although windshields of the HPR (High Penetration Resistance) type in cars in North America are not at all significantly involved in ocular injuries still, about 50% of the injuries of the eye area are caused by glass. The HPR windshield probably is the main reason for the relatively low occurrence of ocular injuries in United States crashes compared to these injuries reported from countries with tempered windshields.

No ocular injuries were observed among belted occupants in this study. Increased use of lap-shoulder belts would decrease the likelihood for occupant contact with the windshield, mirrors, A-Pillar, steering column and instrument panel--about half of the occupant contacts for ocular injury--and thus further reduce the incidence of injuries leading to decreased vision.

Data from the first fifteen months of the NCSS program provide an estimate of 7.5 cases of blindness in one eye among 62,000 occupants of passenger cars damaged severely enough to require towing from the scene. Among this group, no survivors were blinded in both eyes.

6.5 Cervical Injuries

Previous studies due to traffic accidents cervical injuries have been based on clinical reviews, and not from a known automotive accident population. The National Crash Severity Study has provided for the first time an adequate sample of actual crashes so that cervical injury frequencies and severities can be determined.

The NCSS dataset used for this study represents 62,026 occupants of towed passenger cars, and thus the frequencies quoted here apply to such a group. Of all such persons, one in 300 had a cervical injury in the range AIS-3 to AIS-5 or fatal. For contained occupants (i. e., persons who were not ejected from their cars) this rate was one in 433. For ejected occupants, the rate was one in fourteen. There are, of course, many differences between accidents in which persons are ejected as compared with those in which persons were not ejected, but this one factor seems to be the strongest explanation of serious neck injury.

Given an AIS 3 to 5 level neck injury to an occupant, the worst non-neck injury to that occupant is less severe than the neck injury in 62% of the cases.

More occupants sustain severe neck injuries in frontal or side impacts, but the rate of such injuries is higher in rollovers than in any other crash type. Such injuries are relatively rare in rear-impacted passenger cars.

Car occupants between 16 and 25 years of age sustain severe to fatal neck injuries four times as often (0.43%) as those younger than 16 and twice as often as do those older than 25 years. Of the 131 more severe neck injuries listed for the 130 car occupants, 53 (40%) led to a fatality; all these injuries were in the cervical spine. There were 8 individuals who had injuries of level AIS-3 or AIS-4 in the anterior neck. Most of these involved throat structures, including fractures or transection of the larynx or trachea, or lacerations of the neck involving major blood vessels or their branches.

Rarely, if at all, is the neck fractured or dislocated by direct impact to the cervical area. The anterior neck structures, however, are

almost always injured by direct blunt impacts or impacts causing deep lacerations .

For those not ejected from the car the more serious or fatal neck injuries are more often associated with windshield contacts, although many of these may also involve contact with more substantial structures around the windshield.

6.6 Comments on the NCSS Data

In the process of reviewing the field notes and photography for these clinical studies, the clinical analysis staff has had an opportunity to observe the quality and peculiarities of the original reports. Case selection for detailed reading was generally accomplished by using the computer file for sorting and identification, and then selecting and reading the cases from the files maintained at CALSPAN or at the MVMA.

A first problem in this process occurs when the computer file does not contain the detail necessary to case identification. Most of the clinical studies addressed particular injury types or body regions, and when the Occupant Injury Classification codes were not assigned, some cases of interest would be missed. For example, nearly half of the fatalities in the NCSS data had no OIC's coded, since there was not an autopsy or medical examiner's report provided. This deficiency was largely countered by a full review of all fatal cases, assigning AIS and OIC values (admittedly with lower confidence). We would recommend that future data collection procedures provide for (computer) recording of as much injury detail as possible, with appropriate notation as to the quality of the source. Our review of such (fatal) cases was necessarily performed with available written material, and it seems likely that the original investigators could have more information at hand to make such estimates. A similar argument would hold for other coded variables--including the Collision Deformation Classification (CDC) and Delta V.

A specific example will illustrate this problem. If an analyst wished to study cases in which an occupant suffered a closed head injury at AIS-3 or above, he would ask the computer to list all cases with a head injury in that range. But the difference between an AIS-2 and

AIS-3 concussion depends on the length of time the person was unconscious (more or less than 15 minutes). The investigator may be unable to determine a precise value, and is then required to enter an "unknown" value for the severity level--even though he is certain that it must be either a "2" or a "3". Some provision should be made for recording the degree of uncertainty, perhaps with a most probable value and a range.

Reviewing of the field records was made more difficult than it might have been because of variability in the reporting forms from time to time and team to team, penciled entries and marginal notes which sometimes bordered on illegibility, and the lack of a simple narrative describing the general sequence of events in the crash.

The reporting form variability was undoubtedly the result of a fast startup for the NCSS program. Several teams collected injury data, for example, on forms they had used in previous studies. As a result, one team's pink form would be the equivalent of another's blue form. While this may not have constituted a serious problem for keypunching, it often frustrated the clinical reviewers.

It seems likely that the field personnel made marginal notes for their own use. Yet it is just such notation which may provide a necessary piece of information for a clinical reviewer. In a specific case, a contact for a head injury was coded as "unknown", but a scribbled marginal note (which took some effort to decipher) said something like, "Mr. xxxx stated that his head hit ----- (pointing to the A-pillar)." It may have been quite appropriate, given the ground rules of the study, to code the "unknown" value, but, for a specific research task, the marginal information was most helpful.

The clinical reviewer usually wants to form a general picture of the accident circumstance, learning, for example, that two cars met at a certain angle or speed, had more than one impact, etc. Such information may be derivable from coded entries--for example, the presence of two Collision Deformation Classification codes would suggest a second impact, but this is not easy to infer from the available material.

If the occupant kinematics, as they are understood by the field investigator, were included in a narrative of the accident, this would lead to a more accurate picture of the injury sequence than if the reviewer had to construct the picture from a sequence of Occupant Injury Classification codes, impact values, etc. Further, the process of writing the narrative may assist the field investigator in his coding tasks, and result in a more complete and accurate report.

6.7 Recommendations

Some specific recommendations are given here as possible guidance to future data collection programs.

1. Push the field investigators to make estimates of the injury details, including reporting them in OIC form for computerization. There were many cases in which relatively accurate injury information could be inferred from information available to the investigators, and there should be a system for recording the best detail possible.

2. Contacts which were responsible for various injuries should be better and more completely reported. Some of the deficiency here seems to be the result of minimal training of investigators, but, again, the field people should be encouraged to make best guesses and tag them as such.

3. Field investigators should be informed that analysts and clinical reviewers may be reading their field notes. They should complete their reports with such reviewers in mind, and try to make their marginal notes understandable to all.

4. The procedure in the NCSS program of completing essentially only the coding forms leaves much to be desired. The clinical reviewer would be aided materially by a short (one or two page) summary of the accident and occupant kinematics sequence. This shortcoming was particularly noticeable in the NCSS cases because of the variability of reporting forms, but would be helpful also in more structured recording.

5. Photographic quality in many of the NCSS cases was marginal. Some of this evidently resulted from difficulties in getting inside the vehicles, but some of it from limited photographic training of the investigators. Given the total cost of a NCSS investigation, it would seem appropriate to spend a few more dollars per case to insure that all photos are usable.

7 IMPLICATIONS FOR NASS

The analyses used in this report focused on two different approaches to analyzing the NCSS data. The sample design was used as a basis for obtaining probability-based estimates of the accident experience for the aggregate of the seven NCSS areas. On the other hand, the accident analysis models require analysis of relationships between variables within a specific subset of accidents. Here mechanistic models were developed to predict injury severity from crash severity. These two approaches to analysis are not only technically different, but address different questions. The relationship between these two approaches is discussed and compared by Brewer and Mellor[66] and Holt and Smith.[67]

The purpose of this section is to indicate implications from the analyses done with the NCSS data for analysis of the National Accident Sampling System[68] (NASS) data. NASS is an ongoing probability sample of accidents within the United States sponsored by NHTSA. The primary selection units are a stratified sample of sites. Within each site, a probability sample of accidents is chosen. The NASS design differs from the NCSS design in that the primary selection sites are chosen randomly within strata; the NCSS sites were purposively chosen. Another difference is that the weights attached to particular accidents will in general be more complicated than the weights assigned to the NCSS accidents since NASS investigates all traffic accidents including pedestrians, motorcycles and heavy trucks. Finally when NASS is fully

---

[66] K. R. W. Brewer and R. W. Mellor, "The Effect of Sample Structure on Analytical Surveys," The Australian Journal of Statistics, 15:3 (1973), 145-152.

[67] D. Holt and T. M. F. Smith, "The Design Surveys for Planning Purposes," The Australian Journal of Statistics, 18:1-2 (August 1976), 37-44.

[68] H. John Edmonds, Robert H. Hanson, David R. Morganstein, Joseph Waksberg, National Accident Sampling System Sample Design, Phases 2 and 3--Volume 1: Final Technical Report, Report No. DOT-HS-805-274. Sponsored by the National Highway Traffic Safety Administration, Department of Transportation, Washington D. C., under contract No. DOT-HS-7-01706 (Springfield, Va.: National Technical Information Service, November 1979).

operational it will have 75 primary sites in the United States; NCSS had data collection in seven sites.

In the following, further research and analysis is discussed in five areas. These areas parallel work done with the NCSS data. Extensions of the modelling effort to develop mechanistic models is discussed in Section 7.1. Implications for future work with accident analysis models is presented in Section 7.2. Development of national estimates and related problems are reviewed in Section 7.3. An overview of modelling national statistics is presented in Section 7.4, and possible applications are presented. Finally, a crucial problem in all analyses done was that of missing or incomplete data. This topic is discussed in Section 7.5, and recommendations for further research are presented.

7.1 Mechanistic Models

The objective of this analysis was to determine models appropriate for predicting injury severity from crash severity. In this analysis a recoded variable that dichotomized injury severity was used as the dependent variable. The logit model was chosen to be used in this analysis. The accident analysis models require a distribution for OAIS as a function of crash severity. The model used in this project has been generalized to model, as the dependent variable, an ordinal variable. This technique is described by Aitchison and Silvey.[69] Investigation into whether this model is appropriate for OAIS would be a logical continuation of this analysis.

All of the analysis done in this project was done within collision types. These classes were defined by area of the vehicle damaged, direction of impact, and seat position. These types were an attempt to form groups of accidents homogeneous within groups and different between groups. The models would then reflect differences in the subsets. Another approach would be to focus on injury to a generalized body region and use this to classify occupants into groups. This has the

---

[69] J. Aitchison and S. D. Silvey, "The Generalization of Probit Analysis to the Case of Multiple Responses," Biometrika, 44 (1957), 131-148.

advantage of making the OAIS more meaningful within types. To approach the problem in this manner would require information coded on the level of injury for each generalized body region, including no injury. There is not enough information available in NCSS to do this since only up to six injuries are coded. Using this approach would involve analyzing multivariate dependent variables, each with ordinal categories. This type of analysis would need some further research.

Our work with the mechanistic models has also generated some suggestions with regard to the quantification of the dependent variable used, and also some suggested additions to the independent variables available. The dependent variable used was a dichotomy based on the Abbreviated Injury Scale. Mispredictions were found to frequently involve relatively low to moderate collision severity impacts which resulted in severe injuries. For example, a substantial group of mispredictions in the Near PCD subset of the side impacts (passenger compartment damage, near-side occupants) were ankle dislocations. These receive an AIS 3 code. Many factors were assimilated in developing the AIS scale: threat-to-life, treatment period, probability of permanent impairment, etc. Not all these factors are directly related to the collision forces, especially when comparisons involve different body regions and/or injury types. The ankle dislocations do not seem to be the result of greater collision forces, but rather the particular point and, perhaps, direction of force application. The prediction of this sort of injury would seem to be beyond the capability of models based on variables which define the collision configuration, no matter how detailed the description.

Many of the outliers, or mispredictions, from the mechanistic models developed seemed to have similar explanations. Ruptures and hemorrhages were other severe injuries which were frequently incorrectly predicted in the low to moderate collision severity range by the mechanistic models. The more general problem involved here seems to be the comparison between injuries to different body regions and of different types. This problem is critical for the accident analysis models because our models indicate that this problem is a large part of the reason that only about 50% of the severe injuries are predicted

correctly. The implication of this finding is that a substantial proportion of the severe injuries are not strongly linked to collision severity, and, therefore, the estimated benefits projected by an accident analysis model which presumes all injuries to be "caused" by collision severity will be substantially overestimated. The alternative is to develop more sophisticated dependent variables and associated models.

One approach might be to extend the injury scale development to encompass several different scales for each of the various dimensions such as threat-to-life, treatment period (or even cost), probability of permanent impairment, force required, energy required, etc. Injuries could then be coded on each of these scales, and one would have a multidimensional dependent variable to work with.

At this point, subsetting on body region in addition to collision type would seem to be the more feasible approach despite the uncertainty of the NCSS data. Separate models would then be developed for each generalized body region (say head and neck, torso, and extremities). For use in the accident analysis models, the national accident experience would also have to be characterized by the frequency of injury to each of these generalized body regions. Since all injuries were not coded in the NCSS, there would be some uncertainty as to whether a given body region in fact had no injury, or that the injury was not coded injuries, this problem might not be serious.

Improvements could also be made in the independent variable, collision severity. Currently, Delta V is the primary variable in the data file. The current accident analysis models are written in terms of closing speed so that the effect of altered weight distributions in future vehicle populations can be treated. While this variable can be generated after the fact, our work has indicated that there are some problems in determining which two vehicles were used in the original CRASH run. It would seem simpler if this information were coded in the first place. The clinical analysis of the side impact cases suggested that angular velocity may play an important role in determining the occupant contact point in non-passenger compartment impacts. A first approximation to the post-impact angular velocity can be added to the

existing CRASH algorithm with a single statement, and would be useful if output and coded. Other useful outputs would be the ratio of peak forces for the two vehicles, and the energy absorbed by each. Again, these quantities are currently generated by the program, but simply not output. In fact, the energy absorbed quantity could be computed even when no information was available on the other vehicle, and might be very useful in addressing the current 40% missing data rate on Delta V.

7.2 Accident Analysis Models

The need for national estimates of the accident experience in the accident analysis models was discussed in Section 5. The NCSS data do not provide probability-based estimates of the national experience. However, the National Accident Sampling System will. The NASS statistics will also provide estimates of their variance. At that time, it will be appropriate to modify the accident analysis models to carry these variances through to the projected benefits. In this way the influence of the accuracy of the accident data will be reflected in the estimated benefits. This information should assist in the interpretation of the results of the accident analysis models.

Another important issue concerns the possibility of year-to-year changes, or trends in the national accident experience. Currently, the KRAESP model uses a single set of accident statistics. While the total number of accidents is adjusted to reflect estimated changes in total vehicle use, no changes are made to any of the distributions. If vehicle populations are hypothesized with appreciably different distributions of car size, for example, then it might also be reasonable to envision that the use of these vehicles might be different. Smaller cars tend to be used more in urban environments. The accident experience of vehicles in urban usage is appreciably different from those in rural usage. Collision severity is lower for urban accidents, as is the proportion of single-vehicle accidents. Other trends unrelated to the vehicle population might also be important. These might arise due factors such as oil shortages or changes in national speed limits.

Such trends could be treated analytically with population models like those which were developed to provide national projections from the

NCSS data. This topic is discussed further in Section 7.4, Modelling Population Statistics.

Adjusting the distributions in the national accident experience to reflect year-to-year changes or trends is a subset of the larger problem of accident causation. Although the accident analysis models were originally developed to address the crashworthiness area, the KRAESP model has been extended to make projections in the accident causation area. The user has the option to provide inputs on the type of brake system for each vehicle. Each system is, in turn, presumed to modify the distribution of closing speeds in the accident population. While this seems a plausible first approximation, it opens up a vast and relatively unknown area. It has already been stressed that the validity of the crashworthiness projections rests on the degree to which the mechanistic models reflect the governing physical principles. Problems previously discussed in the crashworthiness area include collision types for which good measures of collision severity do not exist such as rollovers, and substantial portions of the severe injuries that do not appear to be well determined by the collision severity.

Similarly, projections in the causation area will only be as good as the models used to make the projections. The validity of the accident causation models will also rest on their foundation in the physical principles which determine the occurrence or non-occurrence of an accident. This treatment of the accident causation process is felt to be far more difficult then the crashworthiness area. Unfortunately, assumptions in this area are currently required in the accident analysis models just to adjust the national accident experience to reflect changes in the total estimated vehicle use resulting from changes in the vehicle population as already mentioned. The total number of accidents is adjusted to reflect estimated changes in total vehicle mileage. No differences in vehicle mileage are included for vehicles of differing sizes. Only the total number of accidents is adjusted, so that the distributions remain unchanged. These assumptions imply an accident causation model. It will be important to review the implications of these assumptions.

7.3 Descriptive Population Statistics

An important use of the NCSS data was to provide summary statistics that describe the accident population studied. These statistics from NCSS, without adjustment, may not be adequate to describe the U.S. accident population. NASS will provide probability-based estimates for the U.S. accident population. The estimation procedures are specific to the NASS design and are discussed in the final technical report on NASS.[70]

In this report the method of calculating sampling errors is also presented. In designing software to calculate estimates, the sampling error should also be calculated. It is very difficult however to tabulate large numbers of statistics and along with each its sampling error. Design effects[71] were calculated for selected NCSS statistics and some patterns could be seen. These design effects will not be applicable to NASS since NASS is based on a completely different sample design. Investigation of design effects to summarize sampling errors needs to be done over a period of time to see if consistent groups of variables appear with the same design effects.

To adjust the NCSS data to reflect the national accident experience a method was developed to produce a national projection. This method is not restricted to the NCSS data. The projection method could be thought of as a type of post-stratification. It is possible that by using the projection method a better estimate would be obtained. This is particularly true when for a certain variable the original stratification of the NASS design is not efficient and there is known to be a better stratification. It is this stratification variable that the projection method would be based on.

---

[70] H. John Edmonds, Robert H. Hanson, David R. Morganstein, Joseph Waksberg, National Accident Sampling System Sample Design, Phases 2 and 3--Volume 1: Final Technical Report, Report No. DOT-HS-805-274. Sponsored by the National Highway Traffic Safety Administration, Department of Transportation, Washington D. C., under contract No. DOT-HS-7-01706 (Springfield, Va.: National Technical Information Service, November 1979).

[71] Leslie Kish, Survey Sampling (New York: John Wiley & Sons, 1965).

7.4 Modelling Population Statistics

The analysis of complex survey data has been recently given some attention in the statistical literature. These "population models" involve modelling population statistics obtained from survey data. In contrast, the "mechanistic models" are models that describe particular elements in the population. If, in fact, all elements were described by the same mechanistic model there would be no difference between the population and mechanistic models. But if the population is made up of subsets that are described by different mechanistic models the overall population model will describe "an average" of the individual mechanistic models.

These population models could be simple models such as independence in a contingency table, or more complicated models such as regression analysis, time series, or the logit analysis. Koch, Freeman, and Freeman[72] present a general framework for using the weighted least squares approach in analyzing these population models. This is an extension of work presented by Freeman[73] where regression analysis is discussed in this context. Shuster and Downing[74] present a method to test for independence when the proportions in the contingency table are population estimates based on a sample design. Rao and Scott[75] have been investigating goodness of fit models and models of independence for contingency tables based on complex survey data. They have found that chi-square tests for goodness of fit need to be adjusted by design

---

[72]Gary G. Koch, Daniel H. Freeman, Jr., and Jean L. Freeman, "Strategies in the Multivariate Analysis of Data from Complex Surveys," International Statistical Review, 43:1 (1975), 59-78.

[73]D. H. Freeman, The Regression Analysis of Data from Complex Sample Surveys: An Empirical Investigation of Covariance Matrix Estimation, mimeo ser. No. 1020 (Chapel Hill, N. C.: Institute of Statistics, 1975). 1975.

[74]J. J. Shuster and D.J. Downing, "Two-way Contingency Tables for Complex Sampling Schemes," Biometrika 63:2 (1976), 271-276.

[75]J. N. K. Rao and A. J. Scott, "The Analysis of Categorical Data from Complex Sample Surveys I: Chi-squared Tests for Goodness of fit," Paper presented at the American Statistical Association Meeting, August 1979.

effects or generalized design effects. Finally, Smith[76] considers problems of time series analysis when there is an on-going survey.

One possible avenue of research is to see whether these methods are useful in analyzing the NASS data, in accident analysis models, in developing causal models, and in analyzing time series. Causation models may well be expressed in terms of these population models and incorporated into accident analysis models to allow for a more realistic picture of the accident environment. After three years of data collection in NASS there will be enough data to estimate monthly time series.

7.5 Incomplete Data

While working with the NCSS data, missing data appeared to be one of the biggest obstacles. There can be no doubt that the best solution to the missing data problem is to encourage field investigators to employ procedures to minimize missing data. But missing data continues to be a problem and a statistical approach to the problem of missing data is also necessary.

In the NCSS Statistics publications, distributions based on variables with missing data were calculated including a separate category for missing data. The analysis done in Section 4.5 indicated that distributions of key NCSS variables differs in the cases with missing data, so that distributions from NCSS ignoring missing data may be biased.

In the development of the mechanistic models, occupants with missing data on any variable included in the model had to be excluded from the logit analyses. The effect is that only a relatively small subset of occupants are included. This is not a problem specific to the logit analysis, but would occur in any multivariate analysis unless adaptive procedures are developed.

These two analytical problems are similar in that the focus is on contingency tables: distributions being a contingency table with only

---

[76] T. M. F. Smith, "Principles and Problems in the Analysis of Repeated Surveys," in Survey Sampling and Measurement, ed. N. Krishnan Namboodiri (New York: Academic Press, Inc., 1978), 201-216.

one dimension. The analysis techniques differ by whether or not the sampling weights are included. Weighted data to generate the distributions in NCSS Statistics are used to obtain unbiased (or approximately unbiased) estimates of aggregate totals and proportions. In model development the most important concern is to obtain good estimates of specified model parameters to best describe the mechanistic model. Analysis modifications will be heavily dependent on the type of analysis, whether estimation of distributions, modelling of contingency tables, or a logit analyses. The effects of weighting will be of secondary importance.

One approach to the missing data problem is to impute for each case with a missing item a "best predicted value." The advantage of this approach is that the data set is now "complete." Tabulations produced from such a data set will always give consistent totals. These procedures use information obtained from the non-missing cases in the sample to obtain the predicted value.

Many procedures have been devised for on-going surveys where missing data occurs. A traditional approach is to use a variation of the "hot deck" procedure.[77] But there are alternative imputation procedures. The EM algorithm[78] can be used in this context. A general procedure referred to as "weighting class adjustments"[79] is another alternative. All of these methods can be used to modify a data set so that it resembles a "complete" data set. With the many available procedures, the final choice depends on the type of data and the use to which the data is intended.

---

[77] Innis G. Sande, "Hot Deck Imputation Procedures," Symposium on Incomplete Data: Preliminary Proceedings (Washington, D. C., Social Security Administration, December, 1979), pp. 484-507.

[78] A. P. Demster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithum," Journal of the Royal Statistical Society, ser. B, 39:1 (1977), 1-38.

[79] David W. Chapman, "A Survey of Nonresponse Imputation Procedures," American Statistical Association Proceedings of the Social Statistics Section, 1976: Part I (Washington, D. C.: American Statistical Association, 1976), pp. 245-251.

Every discipline has its own variables unique to that area. Imputation methods need to be designed with the nature of the variable in mind. Accident variables may be different than agricultural or economic data. Different procedures may be more appropriate for categorical or ordinal or continuous variables.

There are various criteria that need to be evaluated in choosing an imputation procedure. All of these imputation procedures affect the bias, variance, and correlation structure of the resulting estimates. Care must be taken in the choice of the procedure. Ideally the bias should be minimized and variances and correlations affected as little as possible. The choice of a particular method becomes complex when many variables are involved, so that key variables need to be defined and procedures chosen with respect to these variables. Finally, there are different costs associated with each imputation procedure. These costs may involve computer time and development time. Ultimately the best procedure will be a compromise between costs and the effect of the imputation procedure on the data.

The other approach to missing data is to modify the statistical methodology to incorporate more of the data available in the analysis. There has been some work in this area in regression analysis and categorical analysis. Haitovsky[80] looks at the effect of missing data in regression analysis. Various methods for using more of the available data in regression analyses are discussed and evaluated. Little[81] investigates the effect of the different methods on statistical tests and confidence intervals for the regression coefficients.

When analyzing categorical data, different models may be assumed. Hocking and Oxspring[82] consider multinomial sampling where some

---

[80] Yoel Haitovsky, "Missing Data in Regression Analysis," Journal of the Royal Statistical Society, ser. B, 30:1 67-82.

[81] Roderick J. A. Little, "Maximum Likelihood Inference for Multiple Regression with Missing Values: A Simulation Study," Journal of the Royal Statistical Society, ser. B, 41:1 (1979), 76-87.

[82] R. R. Hocking and H. H. Oxspring, "Maximum Likelihood Estimation with Incomplete Multinomial Data," Journal of the American Statistical Association, 66:333 (March 1971), 65-70.

observations have missing data on one or more variables. They derive maximum likelihood estimates for the multinomial distribution using variables with only marginal distributions. Adaptation of the weighted-least squares approach to analyzing contingency tables where observations have some missing data is presented by Koch, Imrey, and Reinfurt.[83] There is no literature examining the effect of modifications of the analysis for missing data.

Although investigation of the NCSS data does not provide a good indication of missing data rates or types of missing data in NASS, it does point out that the analysis of data becomes much more complex in the presence of missing data. NASS will undoubtedly have a certain amount of missing data. From the experience of analyzing the NCSS data missing data will affect two major areas, the presentation of distributions for accident statistics and a reduction in sample sizes available for developing statistical models.

Future activity in this area would include a review of imputation procedures. For NASS an imputation procedure could be chosen such that the bias is minimized, the covariance structure maintained and is cost effective. Software could be developed to create a "complete" data set to be used for tabulation of accident statistics. Additional work in modification of statistical methods to incorporate more available data could also be undertaken. Modifications such as these will allow more of the available data to be used in the statistical analysis.

---

[83]Gary G. Koch, Peter B. Imrey, and Donald W. Reinfurt, "Linear Model Analysis of Categorical Data with Incomplete Response Vectors," Biometrics, 28 (September 1972), 663-692.

APPENDICES

APPPENDIX A
THE ALGORITHMS FOR CREATING THE NEWOAIS INJURY VARIABLES

APPPENDIX A
THE ALGORITHMS FOR CREATING THE NEWOAIS INJURY VARIABLES

The algorithms for calculating the NEWOAIS injury variables were written at NCSA. They are described in a memorandum from Sue Partyka at NHTSA, dated March 1979 and entitled "Documentation for Newoais injury variables."

The recode is integer.

To calculate NEWOAIS2:

```
IF 'INJURY SEVERITY -POLICE' EQUALS 5 THEN 'NEWOAIS2'=0
IF 'INJURY SEVERITY -POLICE' EQUALS 1 THEN 'NEWOAIS2'=1
IF 'INJURY SEVERITY -POLICE' DOES NOT EQUAL 5 OR 1 THEN
   'NEWOAIS2'=MISSING
IF 'NCSS CLASS' EQUALS 8 THEN 'NEWOAIS2'=0
IF 'NCSS CLASS' EQUALS(1-3) THEN 'NEWOAIS2'=1
IF 'OIC1-AIS SEVERITY' EQUALS(0-1) THEN 'NEWOAIS2'=0
IF 'OIC1-AIS SEVERITY' EQUALS(2-6) OR 'OIC2-AIS SEVERITY' EQUALS(2-6) OR
   'OIC3-AIS SEVERITY' EQUALS(2-6) THEN 'NEWOAIS2'=1
IF 'OAIS' EQUALS(0-1) THEN 'NEWOAIS2'=0
IF 'OAIS' EQUALS(2-6) THEN 'NEWOAIS2'=1
```

To calculate NEWOAIS3:

```
IF 'INJURY SEVERITY -POLICE' EQUALS 5 THEN 'NEWOAIS3'=0
IF 'INJURY SEVERITY -POLICE' EQUALS 1 THEN 'NEWOAIS3'=1
IF 'INJURY SEVERITY -POLICE' DOES NOT EQUAL 5 OR 1 THEN
   'NEWOAIS3'=MISSING
IF 'NCSS CLASS' EQUALS 8 THEN 'NEWOAIS3'=0
IF 'NCSS CLASS' EQUALS(1-3) THEN 'NEWOAIS3'=1
IF 'OIC1-AIS SEVERITY' EQUALS(0-2) THEN 'NEWOAIS3'=0
IF 'OIC1-AIS SEVERITY' EQUALS(3-6) OR 'OIC2-AIS SEVERITY' EQUALS(3-6) OR
   'OIC3-AIS SEVERITY' EQUALS(3-6) THEN 'NEWOAIS3'=1
IF 'OAIS' EQUALS(0-2) THEN 'NEWOAIS3'=0
IF 'OAIS' EQUALS(3-6) THEN 'NEWOAIS3'=1
```

To calculate NEWOAIS4:

```
IF 'NO OF DAYS IN HOSPITAL' EQUALS 0 THEN 'NEWOAIS4'=0
IF 'NO OF DAYS IN HOSPITAL' DOES NOT EQUAL 0 THEN 'NEWOAIS4'=MISSING
IF 'INJURY SEVERITY -POLICE' EQUALS 5 THEN 'NEWOAIS4'=0
IF 'INJURY SEVERITY -POLICE' EQUALS 1 THEN 'NEWOAIS4'=1
IF 'NCSS CLASS' EQUALS 8 THEN 'NEWOAIS4'=0
IF 'NCSS CLASS' EQUALS(1-3) THEN 'NEWOAIS4'=1
IF 'OIC1-AIS SEVERITY' EQUALS(0-3) THEN 'NEWOAIS4'=0
IF 'OIC1-AIS SEVERITY' EQUALS(4-6) OR 'OIC2-AIS SEVERITY' EQUALS(4-6) OR
   'OIC3-AIS SEVERITY' EQUALS(4-6) THEN 'NEWOAIS4'=1
```

```
IF 'OAIS' EQUALS(0-3) THEN 'NEWOAIS4'=0
IF 'OAIS' EQUALS(4-6) THEN 'NEWOAIS4'=1
```

APPENDIX B
DATA STRUCTURE FOR VARIANCE COMPUTATIONS

APPENDIX B
DATA STRUCTURE FOR VARIANCE COMPUTATIONS

In order to calculate the sampling variances for NCSS statistics it is necessary to create a data structure that summarizes data by cluster at the appropriate sampling unit, whether day or accident. The clusters are defined by Julian date. Each design group, excluding HSRI and SWRI, used a different systematic sample of days. This adds a complexity to the problem since information is needed for each specific design group sample design in order to select the appropriate Julian date to represent the cluster in the date structure. It is necessary to have each case, represented by a particular design group and Julian date, contain the total of the variable of interest.

The first step in this task is the creation of a dataset in which each case would represent a design group/Julian date. This dataset would have 4550 cases (10 for the number of design groups, 455 for the number of days on the first fifteen months of the study).

A series of variables then need to be created. A design group variable, numbered 1-10, is created by coding the first 455 cases as 1, the second 455 as 2, etc. A year variable is created by coding the first 365 days for each design group as 7, the last 90 days as 8 (for 1977 and 1978). Similarly, a month variable was created within each design group year combination. First the days ordinalized within each design group year and then for each year the first 31 days were coded 1, the second 28 days 2, the next 31 days 3, etc.

Next within each design-group-year-month combination the days were numbered in sequence. Thus a dataset is created that identified for each design group, every Julian date in the fifteen month period. Two design groups were dropped from the dataset. They were the teams that sampled by accident rather than by day: HSRI and SWRI.

Two indicator variables were created, one to indicate appropriate days within each design group for the 25% systematic sample of days, the other to indicate appropriate days for the 10% systematic sample of days. Each was coded with the appropriate design group number (1-10) for days on which the sample was drawn, and with 0 for days on which the sample was not supposed to be drawn. For the 25% sample indicator every

fourth day from the design-group's starting day was coded with the number of the design-group. For the 10% sample indicator every tenth day from the starting day was coded.

Two ID variables were created, one for the 25% stratum and one for the 10% stratum. A linear combination of the indicator variable (coded with the number of the design group), the year, the month and the day identifies each cluster. By recoding team to a similar ten-level design group and forming the same linear combination the same variables are created in one of the original datasets (accident, vehicle or occupant). When the cases in the original dataset are restricted to the appropriate sampling stratum (using a filter) matches are made between the original dataset and the sample day dataset. This matching provides a way to count the overall number of accidents, vehicles or occupants on each appropriate design-group-day. Counts could also be made for any category of interest, e.g. rural accidents, belted occupants. These counts were then written into two permanent datafiles one for the 25% sample, one for the 10% sample. This was done by writing them out with a filter from the file initially created. The filter was the non-missing cases on the 25% or the 10% indicator variable. Thus, days where the count was zero for a variable of interest would still be included in the dataset.

An analogous data structure was created for the two teams that sampled accidents rather than days. For these two teams vehicles and occupants are clustered by accident. A sample dataset "SAMPLE2" was created by taking the team, year, month, day and sequence numbers from the accident file for the two teams. The variable indicating sampling stratum was also included. A linear combination of team, year, month, day nand sequence number was created as an ID variable in the sample dataset and in the vehicle and occupant files. Counts were obtained using the same matching procedure, taking care once again to restrict cases in the original dataset to the appropriate sampling stratum.

Once again the counts were written out into two permanent datasets. This writing was done with a filter in the sample dataset for the sampling stratum so that only appropriate accidents would be included in the count.

It was now possible to obtain counts and statistics by design group at the original level or at the cluster level for any variable of interest. These were four cluster datasets:

1) The eight design groups at the 25% level.
2) The eight design groups at the 10% level.
3) HSRI and SWRI at the 25% level.
4) HSRI and SWRI at the 10% level.

These datasets could be used to calculate means, totals, variances and covariances at the cluster level for any variable by design group. The numbers so produced could be used to calculate the design effects.

APPENDIX C
VARIANCES AND DESIGN EFFECTS FOR SELECTED DESIGN GROUP STATISTICS

APPENDIX C
VARIANCES AND DESIGN EFFECTS FOR SELECTED DESIGN GROUP STATISTICS

Note: the figures in the tables presented in the form ".00360 -3" are equivalent to ".00360 x 10^{-3}".

TABLE 1

Proportions and Variances at the Cluster Level
Accident Statistics

Statistic	Design Group	Estimated Probability	Estimated Variance
Rural	Calspan	.18811	.45696 -3
	HSRI	.43414	.60080 -3
	Ind. A	.53191	.59220 -3
	Ind. B	.67108	.18558 -2
	Kent. A	.61311	.26009 -2
	Kent. B	.78762	.14289 -2
	Kent. C	.20432	.51872 -3
	Miami	.00000	none
	SWRI	.15607	.11280 -3
	DynSci	.00695	.15884 -4
Rush Hour	Calspan	.28926	.35035 -3
	HSRI	.32077	.56360 -3
	Ind. A	.33865	.74008 -3
	Ind. B	.32893	.17175 -2
	Kent. A	.39741	.19564 -2
	Kent. B	.77671	.47997 -2
	Kent. C	.31063	.10103 -2
	Miami	.32951	.28734 -3
	SWRI	.32002	.22940 -3
	DynSci	.32506	.54152 -3
Dry Road	Calspan	.42552	.20958 -2
	HSRI	.51723	.63090 -3
	Ind. A	.58067	.25299 -2
	Ind. B	.57849	.28678 -2
	Kent. A	.50761	.70669 -2
	Kent. B	.53399	.44239 -2
	Kent. C	.61574	.23915 -2
	Miami	.76201	.29191 -2
	SWRI	.80514	.16320 -3
	DynSci	.87445	.79353 -3

TABLE 2

Proportions and Variances at the Cluster Level
Vehicle Statistics

Statistic	Design Group	Estimated Probability	Estimated Variance
Front CDC	Calspan	.56496	.28289 -3
	HSRI	.50752	.39490 -3
	Ind. A	.60863	.46035 -3
	Ind. B	.49020	.03004 -2
	Kent. A	.40560	.20959 -2
	Kent. B	.45635	.20708 -2
	Kent. C	.54397	.87574 -3
	Miami	.40634	.20048 -3
	SWRI	.55576	.07030 -3
	DynSci	.38770	.43309 -3
Right CDC	Calspan	.96884 -1	.88485 -4
	HSRI	.02462	.20304 -3
	Ind. A	.00960	.07652 -3
	Ind. B	.04897	.40849 -3
	Kent. A	.79028 -1	.46039 -3
	Kent. B	.90608 -1	.80756 -3
	Kent. C	.00559	.30422 -3
	Miami	.00096	.84078 -4
	SWRI	.02020	.79905 -4
	DynSci	.80000 -1	.02498 -3
Back CDC	Calspan	.50763 -1	.00098 -3
	HSRI	.32699 -1	.55395 -4
	Ind. A	.36202 -1	.00628 -3
	Ind. B	.67225 -1	.00377 -2
	Kent. A	.40977 -1	.24335 -3
	Kent. B	.38204 -1	.25454 -3
	Kent. C	.56209 -1	.08227 -3
	Miami	.37935 -1	.35300 -4
	SWRI	.40330 -1	.30524 -4
	DynSci	.55769 -1	.37092 -4
Intruded	Calspan	.20484	.05062 -3
	HSRI	.23044	.27455 -3
	Ind. A	.29298	.37420 -3
	Ind. B	.32773	.02850 -2
	Kent. A	.20070	.00275 -2
	Kent. B	.27084	.03742 -2
	Kent. C	.08495	.38595 -3
	Miami	.00209	.72026 -4
	SWRI	.20078	.99766 -4
	DynSci	.00346	.00503 -3

TABLE 2 (CONTINUED)

Statistic	Design Group	Estimated Probability	Estimated Variance
Not Intruded	Calspan	.68704	.26036 -3
	HSRI	.66892	.40976 -3
	Ind. A	.66386	.35527 -3
	Ind. B	.63788	.03660 -2
	Kent. A	.53269	.20605 -2
	Kent. B	.50093	.20096 -2
	Kent. C	.73660	.77829 -3
	Miami	.58709	.34003 -3
	SWRI	.66545	.06532 -3
	DynSci	.50807	.50603 -3
Low Delta V	Calspan	.26034	.35248 -3
	HSRI	.25000	.52640 -3
	Ind. A	.25933	.68450 -3
	Ind. B	.24980	.02247 -2
	Kent. A	.20070	.02650 -2
	Kent. B	.08023	.00383 -2
	Kent. C	.38395	.98029 -3
	Miami	.05873	.20297 -3
	SWRI	.29360	.22954 -3
	DynSci	.04605	.20336 -3
High Delta V	Calspan	.29304	.27859 -3
	HSRI	.08502	.30409 -3
	Ind. A	.38734	.66864 -3
	Ind. B	.33384	.05029 -2
	Kent. A	.20463	.00572 -2
	Kent. B	.08232	.02268 -2
	Kent. C	.08802	.54200 -3
	Miami	.04007	.00898 -3
	SWRI	.26070	.07499 -3
	DynSci	.05423	.20535 -3
Hit a Car	Calspan	.53993	.42982 -3
	HSRI	.47632	.60577 -3
	Ind. A	.45575	.58275 -3
	Ind. B	.40489	.27475 -2
	Kent. A	.45560	.09030 -2
	Kent. B	.33080	.07007 -2
	Kent. C	.59066	.99294 -3
	Miami	.57830	.25295 -3
	SWRI	.57262	.23449 -3
	DynSci	.63654	.42508 -3

TABLE 3

Proportions and Variances at the Cluster Level
Occupant Statistics

Statistic	Design Group	Estimated Probability	Estimated Variance
Aged 16 & Under	Calspan	.12947	.19050 -3
	HSRI	.13503	.29898 -3
	Ind. A	.20749	.38462 -3
	Ind. B	.16972	.72879 -3
	Kent. A	.18546	.60692 -3
	Kent. B	.17825	.60604 -3
	Kent. C	.16472	.69061 -3
	Miami	.93454 -1	.71790 -4
	SWRI	.18690	.11175 -3
	DynSci	.84374 -1	.20452 -3
Aged 17 to 30	Calspan	.52055	.33537 -3
	HSRI	.56563	.55651 -3
	Ind. A	.50389	.67464 -3
	Ind. B	.54231	.14782 -2
	Kent. A	.51423	.72747 -3
	Kent. B	.52843	.20176 -2
	Kent. C	.53355	.81057 -3
	Miami	.42785	.17598 -3
	SWRI	.50475	.19077 -3
	DynSci	.48352	.41254 -3
Aged 31 to 45	Calspan	.14942	.13899 -3
	HSRI	.12921	.15816 -3
	Ind. A	.10078	.16075 -3
	Ind. B	.13099	.35020 -3
	Kent. A	.14173	.37366 -3
	Kent. B	.11947	.36842 -3
	Kent. C	.15501	.19871 -3
	Miami	.20312	.10711 -3
	SWRI	.14287	.65790 -4
	DynSci	.19321	.23546 -3
Aged 46 & Over	Calspan	.18202	.16569 -3
	HSRI	.14290	.26384 -3
	Ind. A	.17139	.27421 -3
	Ind. B	.14169	.55056 -3
	Kent. A	.15069	.70200 -3
	Kent. B	.17383	.13614 -2
	Kent. C	.14359	.35838 -3
	Miami	.26362	.19279 -3
	SWRI	.15085	.81783 -4
	DynSci	.15377	.23457 -3

TABLE 3 (CONTINUED)

Statistic	Design Group	Estimated Probability	Estimated Variance
Unbelted	Calspan	.77423	.32791 -3
	HSRI	.69022	.40267 -3
	Ind. A	.83250	.42691 -3
	Ind. B	.85118	.34652 -3
	Kent. A	.69021	.10686 -2
	Kent. B	.79328	.20229 -2
	Kent. C	.80475	.52526 -3
	Miami	.71043	.55492 -3
	SWRI	.76129	.15344 -3
	DynSci	.50051	.48119 -3
Belted	Calspan	.10448	.89736 -4
	HSRI	.10235	.15329 -3
	Ind. A	.44333 -1	.13573 -3
	Ind. B	.30582 -1	.12141 -3
	Kent. A	.30032 -1	.13750 -3
	Kent. B	.35398 -1	.24275 -3
	Kent. C	.69941 -1	.17706 -3
	Miami	.27637 -1	.24831 -4
	SWRI	.87749 -1	.51717 -4
	DynSci	.82127 -1	.13084 -3
With OAIS 0-2	Calspan	.86746	.94289 -4
	HSRI	.75115	.27069 -3
	Ind. A	.86106	.19214 -3
	Ind. B	.81091	.49249 -3
	Kent. A	.92150	.25981 -3
	Kent. B	.92984	.12569 -3
	Kent. C	.90294	.17445 -3
	Miami	.74877	.14602 -3
	SWRI	.80958	.83581 -4
	DynSci	.56590	.41903 -3
With OAIS 3-6	Calspan	.30248 -1	.66904 -5
	HSRI	.34057 -1	.42184 -5
	Ind. A	.34735 -1	.57009 -5
	Ind. B	.74414 -1	.77600 -4
	Kent. A	.46364 -1	.67395 -4
	Kent. B	.39189 -1	.29258 -4
	Kent. C	.26834 -1	.62265 -5
	Miami	.15147 -1	.16575 -5
	SWRI	.32501 -1	.42402 -5
	DynSci	.15727 -1	.34794 -5

TABLE 4

Design Effects
Accident Statistics

Statistic	Design Group	Design Effect
Rural	Calspan	2.3724
	HSRI	1.5683
	Ind. A	1.4903
	Ind. B	3.3195
	Kent. A	2.4993
	Kent. B	1.4201
	Kent. C	.96963
	Miami	none
	SWRI	1.0178
	DynSci	1.8870
Rush Hour	Calspan	1.7177
	HSRI	1.7951
	Ind. A	2.0445
	Ind. B	2.5630
	Kent. A	1.9823
	Kent. B	4.1379
	Kent. C	1.8391
	Miami	1.4856
	SWRI	1.8373
	DynSci	1.5748
Dry Road	Calspan	7.9128
	HSRI	1.7835
	Ind. A	6.4672
	Ind. B	3.9439
	Kent. A	6.5497
	Kent. B	3.3611
	Kent. C	3.7489
	Miami	21.142
	SWRI	1.7334
	DynSci	4.0204

TABLE 5

Design Effects
Vehicle Statistics

Statistic	Design Group	Design Effect
Front CDC	Calspan	1.4226
	HSRI	1.3322
	Ind. A	1.4441
	Ind. B	1.9619
	Kent. A	2.3791
	Kent. B	1.9106
	Kent. C	1.6521
	Miami	1.2759
	SWRI	1.4695
	DynSci	1.4038
Right CDC	Calspan	1.0601
	HSRI	1.5415
	Ind. A	1.1553
	Ind. B	1.2784
	Kent. A	1.7058
	Kent. B	1.9054
	Kent. C	1.2147
	Miami	1.3321
	SWRI	1.5836
	DynSci	1.2181
Back CDC	Calspan	3.2497
	HSRI	1.4389
	Ind. A	2.4511
	Ind. B	11.296
	Kent. A	1.5191
	Kent. B	1.6050
	Kent. C	1.5070
	Miami	1.2664
	SWRI	1.8362
	DynSci	1.5778
Intruded	Calspan	.83504
	HSRI	.97994
	Ind. A	1.1274
	Ind. B	1.8967
	Kent. A	1.1594
	Kent. B	1.2276
	Kent. C	.78500
	Miami	.70161
	SWRI	.93514
	DynSci	.58353

TABLE 5 (CONTINUED)

Statistic	Design Group	Design Effect
Not Intruded	Calspan	1.3074
	HSRI	1.4145
	Ind. A	1.0624
	Ind. B	2.0222
	Kent. A	2.3250
	Kent. B	1.9536
	Kent. C	1.4957
	Miami	2.1131
	SWRI	1.4133
	DynSci	1.6271
Low Delta V	Calspan	2.6397
	HSRI	2.8859
	Ind. A	3.7619
	Ind. B	3.6822
	Kent. A	2.6066
	Kent. B	1.7547
	Kent. C	2.1852
	Miami	2.5214
	SWRI	3.1064
	DynSci	1.4236
High Delta V	Calspan	1.4216
	HSRI	1.2731
	Ind. A	1.9957
	Ind. B	2.2892
	Kent. A	1.4862
	Kent. B	1.1889
	Kent. C	1.1670
	Miami	.99316
	SWRI	1.5977
	DynSci	.93367
Hit a Car	Calspan	2.1497
	HSRI	2.0773
	Ind. A	1.7907
	Ind. B	4.3611
	Kent. A	2.0658
	Kent. B	1.6384
	Kent. C	1.8652
	Miami	1.6235
	SWRI	2.0117
	DynSci	1.4218

TABLE 6

Design Effects
Occupant Statistics

Statistic	Design Group	Design Effect
Aged 16 & Under	Calspan	3.3396
	HSRI	3.1142
	Ind. A	3.1526
	Ind. B	2.9402
	Kent. A	2.0505
	Kent. B	1.5541
	Kent. C	3.7724
	Miami	1.8446
	SWRI	2.6752
	DynSci	2.9318
Aged 17 to 30	Calspan	2.7694
	HSRI	3.0095
	Ind. A	3.3495
	Ind. B	3.7509
	Kent. A	1.5443
	Kent. B	3.2519
	Kent. C	2.5227
	Miami	1.8532
	SWRI	2.8326
	DynSci	2.0818
Aged 31 to 45	Calspan	2.2937
	HSRI	1.8679
	Ind. A	1.8264
	Ind. B	1.9743
	Kent. A	1.7674
	Kent. B	1.2141
	Kent. C	1.1330
	Miami	1.7057
	SWRI	2.0635
	DynSci	1.9856
Aged 46 & Over	Calspan	2.2864
	HSRI	2.8059
	Ind. A	2.4039
	Ind. B	2.8174
	Kent. A	3.1268
	Kent. B	3.4248
	Kent. C	2.0969
	Miami	2.4131
	SWRI	2.1814
	DynSci	2.3434

TABLE 6 (CONTINUED)

Statistic	Design Group	Design Effect
Unbelted	Calspan	4.3481
	HSRI	2.8101
	Ind. A	4.0887
	Ind. B	1.6558
	Kent. A	3.2862
	Kent. B	5.0419
	Kent. C	2.6975
	Miami	7.1539
	SWRI	3.4210
	DynSci	2.4355
Belted	Calspan	2.4274
	HSRI	3.1198
	Ind. A	4.9291
	Ind. B	2.7948
	Kent. A	3.2869
	Kent. B	2.8959
	Kent. C	2.5309
	Miami	2.9528
	SWRI	2.9465
	DynSci	2.4659
With OAIS 0-2	Calspan	1.0548
	HSRI	1.5223
	Ind. A	1.1472
	Ind. B	1.3348
	Kent. A	.85297
	Kent. B	.31328
	Kent. C	.90068
	Miami	1.7379
	SWRI	1.4178
	DynSci	2.1221
With OAIS 3-6	Calspan	.13530
	HSRI	.48631
	Ind. A	.65871
	Ind. B	.29878
	Kent. A	.26477
	Kent. B	.97855
	Kent. C	.51434
	Miami	.78091
	SWRI	.15488
	DynSci	.10712