

# REVIEW OF THE MOTOR CARRIER SAFETY STATUS MEASUREMENT SYSTEM (SAFESTAT)

Ken Campbell ([CampbellKL@ORNL.gov](mailto:CampbellKL@ORNL.gov))  
Rick Schmoyer ([SchmoyerRLJR@ORNL.gov](mailto:SchmoyerRLJR@ORNL.gov))  
Ho-Ling Hwang ([HwangHL@ORNL.gov](mailto:HwangHL@ORNL.gov))

Final Report  
October 2004

Prepared for the  
Analysis Division  
Federal Motor Carrier Safety Administration  
U.S. Department of Transportation  
400 Seventh Street, S.W.  
Washington, D.C. 20590  
Under Reimbursable Agreement No. DTFH61-01-Y-30103  
Modification 7

Prepared by the  
Center for Transportation Analysis  
Oak Ridge National Laboratory  
Oak Ridge, Tennessee 37831  
Managed by  
UT-Battelle, LLC  
For the U.S. Department of Energy  
Under Contract No. DE-AC05-00OR22725

## CONTENTS

	<u>Page</u>
LIST OF TABLES .....	iii
LIST OF FIGURES .....	iv
ABBREVIATIONS .....	v
EXECUTIVE SUMMARY .....	vi
1 INTRODUCTION .....	1
2 DATA ISSUES .....	4
2.1 MISSING RECORDS .....	4
2.2 LATE DATA .....	9
2.3 EFFECT OF LATE DATA ON CARRIER RANKINGS .....	12
2.4 THE IMPACT OF MISSING OR LATE DATA .....	16
3 STATISTICAL METHODS .....	18
3.1 REGRESSION MODELS .....	18
3.2 CONFIDENCE INTERVALS FOR CRASH RISK SCORES AND RANKS .....	26
3.3 AN EMPIRICAL BAYES APPROACH .....	29
4 DISCUSSION OF ISSUES .....	34
5 CONCLUSIONS .....	39
REFERENCES .....	41
APPENDIX A: NOTES FROM VOLPE .....	42
APPENDIX B: CONFIDENCE INTERVALS .....	56

## LIST OF TABLES

<u>Table</u>	<u>Page</u>
1	SafeStat Post-Selection Crash Rates .....2
2	Data Quality Problem States Listed by Volpe and UMTRI .....5
3	Changes in SafeStat versions used for the Original and Simulated March 2001 runs .....12
4	Definition of SafeStat Categories .....13
5	Number of Carriers by SafeStat Category .....13
6	Total Weighted Number of Crashes from March 2001 Run.....15
7	Total Weighted Number of Crashes from Simulated March 2001 Run .....15
8	Difference in the Number of Crashes .....16
9	SafeStat Post-Selection Crash Rates with Power Units and Weighted Crashes.....19
10	Post-Selection Crash Rates for Poisson Model Carrier Groups.....20
11	Post-Selection Crash Rates for LCB Model Carrier Groups .....21
12	A Generic Classification Table.....22
13	SafeStat and Regression Classification Rates.....24
14	SafeStat and LCB-Based Classification Rates.....25
15	Crash Rate Estimates with Upper and Lower 95 percent Prediction Bounds.....28
16	Post-Selection Crash Rates for Negative-Binomial Model Carrier Groups .....31
17	Post-Selection Crash Rates for Empirical Bayes Carrier Groups.....31
18	Negative Binomial Classification for At-Risk Carriers .....32
19	Negative Binomial Classification for All Identified Carriers .....32
20	Empirical Bayes Classification for At-Risk Carriers.....32
21	Empirical Bayes Classification for All Identified Carriers.....32

## LIST OF FIGURES

1	Comparison of Trucks in Fatal Crashes in the 2001 TIFA and MCMIS Crash Files.....	7
2	Comparison of Projected and Actual Counts in the 2001 MCMIS Crash File.....	8
3	Distribution of Time Required to Upload Crash Data.....	10
4	Days to Receive 90 Percent of 2001 Crash Data by State.....	11
5	Confidence Bounds for a Hypothetical Score and its Corresponding Rank.....	29

## ABBREVIATIONS

CI	Confidence Interval
CR	Compliance Review
EB	Empirical Bayes
FARS	Fatality Analysis Reporting System
FMCSA	Federal Motor Carrier Safety Administration
GES	General Estimates System
LCB	Lower Confidence Bound
MCMIS	Motor Carrier Management Information System
OIG	U.S. Department of Transportation Office of the Inspector General
ORNL	Oak Ridge National Laboratory
PU	Power Unit
SafeStat	Motor Carrier Safety Status Measurement System
SEA	Safety Evaluation Area
TIFA	Trucks Involved in Fatal Accidents
UMTRI	University of Michigan Transportation Research Institute
UCB	Upper Confidence Bound
VMT	Vehicle Miles Traveled
VNTSC	Volpe National Transportation Systems Center

# REVIEW OF THE MOTOR CARRIER SAFETY STATUS MEASUREMENT SYSTEM

October 2004  
Oak Ridge National Laboratory

## EXECUTIVE SUMMARY

The Motor Carrier Safety Status Measurement System (SafeStat) was developed for the Federal Motor Carrier Safety Administration (FMCSA) by the Volpe National Transportation Systems Center (hereinafter Volpe) in the mid-1990's to measure the relative safety fitness of commercial motor vehicle operators and guide the deployment of resources to focus on carriers posing the greatest safety risk. SafeStat combines information on crashes, roadside inspections, traffic violations and compliance reviews from the previous 30 months to produce an overall SafeStat score for carriers with sufficient safety data. The scores are ranked to identify high risk motor carriers for subsequent compliance reviews and roadside inspections. The SafeStat algorithm has been revised several times to improve the ranking system. SafeStat has become an important tool to support improved motor carrier safety.

Volpe recently updated a 1998 evaluation of the effectiveness of SafeStat.<sup>1, 2</sup> Both the original (1998) and the updated Volpe evaluations confirmed that the SafeStat system successfully identified high-risk carriers. The updated Volpe evaluation used historical data available as of March 2003. In this analysis, Volpe assessed over time the safety fitness of carriers identified by SafeStat as “at-risk” as of March 2001, using data from the previous 30 months (September 1998—March 2001). Based on this assessment, carriers were classified as “At-Risk,” “Other Identified” and “Not Identified” (i.e., not at risk). The subsequent safety performance of these same carriers was then evaluated using an 18 month follow-up period (April 2001—September 2002). The Volpe evaluation found that the carriers initially identified as at-risk by SafeStat, when taken as a group, experienced a 112 percent higher crash rate in the follow-up period, than the carriers not identified as “at-risk” by SafeStat (52.0 versus 24.6 crashes per power unit). In February 2004, the U.S. Department of Transportation Office of the Inspector General (OIG) issued an audit of SafeStat.<sup>3</sup> Both of the Volpe evaluations and the OIG audit identify potential issues with data quality and the algorithm.

The objective of this project is to review the Volpe evaluation. A more detailed analysis might identify shortcomings in the existing system and provide a basis for further improvement. This project addresses two issues: (1) the impact of missing and late crash

---

<sup>1</sup> Madsen, D.G and Wright, D.G. *An Effectiveness Analysis of SafeStat*. TRB paper No. 990448. November 1998.

<sup>2</sup> Volpe National Transportation Systems Center. *SafeStat Effectiveness Study Update*. March 2004.

<sup>3</sup> U.S. Department of Transportation Office of the Inspector General. *Improvements Needed in the Motor Carrier Safety Status Measurement System*. February 2004.

records on the SafeStat ranking, and (2) the potential for statistical methods to improve the effectiveness of SafeStat.

The Oak Ridge National Laboratory (ORNL) reviewed previous analyses and examined the MCMIS Crash file. Underreporting to the Crash file was estimated at about one-third. While state reporting of fatal crashes was nearly complete, only about 80 percent of injury and 50 percent of tow-away crashes were reported. The distribution of time for FMCSA to receive the crash data from the states showed that half the reports were received in about 6 months and 90 percent were received after 16 months. Based on this distribution, about 25 percent of the crash reports would be missing in the 30 months preceding the current date (the SafeStat window).

Since historical data were used for the Volpe evaluation, all late crash reports were available for the 30 month pre-selection period and about 83 percent were available for the post-selection period. Volpe provided the results of the original ranking prepared in March 2001. ORNL assessed the impact of late data by comparing the original ranking from March 2001 with the Volpe simulated ranking based on the complete historical data. ORNL found that the number of crashes in the original ranking was 26 percent lower than in the historical data for the period. The rankings of 81 percent of the carriers were unchanged with the addition of late data. The number of at-risk carriers was increased by 33 percent by the movement of carriers previously ranked lower. However, this increase was offset by the movement of 15 percent of the at-risk carriers to lower rankings, for a net increase in at-risk carriers of 18 percent. Incomplete data resulted in some carriers being ranked at-risk when they would not have been with complete data. Both the timeliness and completeness of MCMIS data are improving so current rankings would not be expected to be impacted as much as the March 2001 ranking. ORNL concluded that missing data patterns were not random and can be expected to introduce bias, particularly with the small sample sizes for the majority of SafeStat measures.

ORNL also used the crash rates from the 18 month follow-up period to produce a new ranking of the carriers. This analysis focused on the safety fitness of individual carriers in each group, rather than on the group's aggregate risk (which had already been examined by Volpe). The ORNL evaluation shows that 90 percent of the carriers identified as "at-risk" by the Volpe SafeStat algorithm did not have a high crash risk in the follow-up period. However, the current SafeStat algorithm is about twice as effective as random selection in identifying high risk carriers (only about 5 percent of all carriers with sufficient data are identified as high risk).

Various statistical models were considered by ORNL to see if they could improve the effectiveness of the current SafeStat algorithm. The statistical models were constrained for purposes of comparison to use the censored and weighted data and to identify the same number of carriers. All of the statistical models were more effective than the Safestat algorithm, although the improvement was modest (30 percent). The use of unweighted and uncensored data may improve the effectiveness of the statistical methods.

ORNL concludes that the Volpe SafeStat algorithm does not adequately address the inherent variability in the scores when identifying high risk carriers. This leads to the selection of some carriers as high risk when their score does not exceed the inherent variability of the SafeStat scores. In fact, about 90 percent of the carriers identified as at-risk in the simulated March 2001 ranking did not have a high crash risk in the post-selection period. Selecting carriers with the highest scores, without addressing the accuracy of the scores, results in the selection of many carriers due to random variations and not any true change in carrier risk. In this situation, random variations would also be expected to return these carriers to expected risk levels in a subsequent observation period. This phenomenon is referred to as “regression to the mean.” Provisions in the Safestat algorithm to assure sufficient information are not adequate.

Statistical methods are available to quantify the variability inherent to the data and correct for regression to the mean. The application of statistical methods can distinguish carriers with significantly elevated safety risk from those with risk levels that do not exceed the variability in the source data. For example, empirical Bayes approaches are now widely accepted for the very similar problem of selecting highway sites for treatment<sup>4</sup>. Statistical models can be used to select coefficients (weights) for the various measures based on the relationship to collision risk in the historical data. This approach would replace expert judgment with objective statistical methods. While statistical methods can quantify random variations, they cannot correct for bias error. Improving the timeliness and completeness of the source data are still essential.

---

<sup>4</sup> Hauer, E., D.W. Harwood, F.M. Council, M.S. Griffith. *The Empirical Bayes Method For Estimating Safety: A Tutorial*. Transportation Research Record 1784, pp. 126-131. National Academies Press, Washington, D.C. 2002.

# 1 INTRODUCTION

The Motor Carrier Safety Status Measurement System (SafeStat) was developed by the Volpe National Transportation Systems Center (VNTSC) and has been used by the Federal Motor Carrier Safety Administration (FMCSA) to address motor carrier safety since March 1997. Volpe has identified data quality issues and the SafeStat algorithm has been revised several times to improve the SafeStat ranking system. The U.S. Department of Transportation Office of the Inspector General (OIG) issued an audit of SafeStat (1) in February 2004. The issues raised by the OIG included the impact of missing and late data on the carrier rankings and the potential for statistical methods to improve the algorithm.

In March 2004, Volpe updated a 1998 evaluation of the effectiveness of SafeStat in identifying high-risk carriers (2,3). Both the original (1998) and the updated Volpe evaluations found that the SafeStat system identifies high-risk carriers. The focus of this project is to review the Volpe evaluation and address some of the issues identified in the OIG report. A more detailed analysis might identify shortcomings in the existing system and provide a basis for further enhancement.

## 1.1 BACKGROUND

The Volpe SafeStat method scores motor carriers in four Safety Evaluation Areas (SEAs) relating to accidents, drivers, vehicles, and safety management. The algorithm uses data from the Motor Carrier Management Information System (MCMIS). The data sources for SafeStat include the Crash file, Census file, roadside inspections, traffic violations and compliance reviews. SafeStat uses 30 months of data to compute various measures and indicators from the source data that are combined into the four SEAs. Indicators are not calculated unless a carrier has sufficient data. For example, the Accident Indicator is not calculated for carriers with only one accident. The SEA scores are then ranked and the worst 25 percent are classed as unacceptable in that SEA. An overall safety risk score is computed as a weighted combination of the unacceptable SEA values for each carrier that has two or more unacceptable SEAs. A carrier is identified as at-risk if the SafeStat score is 225 or more. The primary objective of the classification is to target high-risk carriers for inspections and compliance reviews.

An updated Effectiveness Study was performed by Volpe to evaluate the current SafeStat method (3). The updated evaluation used historical data available as of March 2003. A simulated ranking was prepared as of March 2001, based on the data for the previous 30 months (September 1998—March 2001). This ranking is referred to as a simulated ranking because any additional data for the 30-month period received as of March 2003 was included and because the current version (Version 8.5) of the SafeStat algorithm was used rather than the version in use in March 2001 (Version 8.2). Based on the March 2001 simulated ranking, carriers were grouped in three safety risk categories: “At-Risk (Category A&B),” “Other Identified (Category C)” and “Not Identified.” “All Identified” refers to the combination of categories A–C.

Volpe defines safety risk as the likelihood of having crashes in the near future. The post-selection carrier performance data for the following 18 months, April 2001–September 2002, were used to calculate the aggregate crash risk for each of the carrier groups. Post-selection data were available for 118,757 carriers, of which, according to the prior SafeStat classification, 5,952 were in the “All Identified” category, 3,595 were in the “At-Risk” category, and 112,805 were “Not Identified.” The results of the Volpe evaluation are repeated in the Table 1 below. The overall finding is that the A&B carrier group (at-risk) had a crash risk 112 percent higher in the post-selection period as compared to the group of carriers not identified. Thus, the post-selection crash risk confirms that the group of carriers identified as at-risk in the March 2001 ranking also had a higher crash risk in the 18 months after they were identified. Volpe concludes that this finding demonstrates the continued effectiveness of SafeStat to identify high-risk carriers.

**Table 1**  
**SafeStat Post-Selection Crash Rates**

<b>Carrier Group</b>	<b>Number of Carriers</b>	<b>Weighted Crashes per 1000 PUs</b>	<b>Percent Higher Than Not Identified Group</b>
All Identified (A-C)	5,952	42.5	73%
At-Risk (A&B)	3,595	52.0	112%
Other Identified (C)	2,357	29.4	20%
SafeStat Not Identified	112,805	24.6	0%

*Volpe SafeStat Effectiveness Study Update, March 2004 (3).*

## 1.2 OBJECTIVE

ORNL was asked by FMCSA to review the updated Volpe effectiveness study and, since the purpose of the study is to evaluate SafeStat, to evaluate the SafeStat method itself. Volpe provided ORNL with the 30-month “pre-selection” and 18-month “post-selection” data and results. Volpe also provided the original SafeStat results for March 2001. Additional notes on the evaluation and documentation for the data were provided by Volpe and are included in Appendix A. ORNL also had three years of MCMIS Crash data covering 2001–2003.

The objective of this project is to review the Volpe evaluation and address some of the issues identified in the OIG report. The issues that have been identified both by the Volpe evaluations and the OIG audit are in two areas: (1) the quality (completeness and timeliness) of the source data and (2) the SafeStat algorithm. Among the issues raised, ORNL selected those that were felt to be most important and could be addressed with the data provided within the available time. Data issues are covered in Section 2 and algorithm issues in Section 3.

The essential data issue is whether (or how) data quality impacts the SafeStat ranking. The most significant data issues are with the MCMIS Crash file and include both missing records (crashes) and late records. The OIG report also identifies problems with the traffic violations data, but this issue could not be examined in this study. Missing Crash records and late Crash records are examined separately. Both Volpe and the University of Michigan Transportation Research Institute (UMTRI) have reported on data quality in the MCMIS Crash file. National truck crash data collected by NHTSA and UMTRI provide a basis to assess underreporting by state for fatal crashes and underreporting of injury and tow-away crashes at the national level. State-to-state differences in reporting are the best evidence available of the potential for bias in missing and late data.

The data Volpe prepared for this review provide a unique opportunity to see the impact of late data on the SafeStat rankings. The impact of late data is not reflected in the Volpe effectiveness study since it is a historical study using all data received as of March 2003. Any late data for the pre-selection period, September 1998–March 2001, would have been received and was used in the simulated March 2001 ranking. To assess the impact of late data, Volpe provided the results of the original March 2001 SafeStat ranking that, of course, did not include the late data. ORNL assessed the impact of late data by comparing the two March 2001 rankings. The distribution of time for the states to upload crash records to MCMIS is also addressed in Section 2.

The SafeStat algorithm is addressed in Section 3. In order to evaluate something, it is helpful to have a comparison. The SafeStat algorithm is the product of expert judgment. The logic is sound and reasonable. Judgments have been made as to how much data are sufficient and how to weight the various components. This review makes the assumption that the most likely alternative is to replace expert judgment with statistical methods. SafeStat is a formula that reflects assumptions about the relationships of various measures to crash risk. The same source data could support statistical models that can also relate the source data to crash risk. Thus, the focus of Section 3 is to make a controlled comparison of the ability of SafeStat and several standard statistical methods to predict carrier risk. A standard evaluation tool for statistical methods is to compare the outcome as determined by the method with the actual outcome for each carrier and this approach provides an alternative way to evaluate both SafeStat and the statistical methods.

The results of this work address many of the SafeStat data and algorithm issues. These are discussed in Section 4. Conclusions are provided in Section 5. References (shown in parentheses) are listed at the end, followed by Appendices.

## 2 DATA ISSUES

FMCSA depends on the states to provide essential data on crashes and roadside inspections for the SafeStat system. It takes longer to receive the crash data and the crash data are most important for the SafeStat system, so that is the focus of this section. There are two aspects to this issue: missing reports and late reports. Late reports are particularly critical to the SafeStat system. For most data uses, one can wait until the file is complete. SafeStat rankings, on the other hand, are computed with all data available at the time of the ranking. Thus, late reports will impact the SafeStat rankings. FMCSA has been working with the states to improve the timeliness and completeness of the MCMIS Crash data since the current system was initiated in 1994 and there has been much progress. The goal is to receive crash reports within 90 days and inspection data within 21 days.

Missing crash reports are addressed first in this section, followed by late reports. Volpe provided the results of the original March 2001 SafeStat ranking that, of course, did not include the late data. ORNL assessed the impact of late data by comparing the original March 2001 ranking with the simulated March 2001 ranking that includes all data received for the pre-selection period as of March 2003.

### 2.1 MISSING RECORDS

Volpe provides current information on the quality of state data used in SafeStat (4). The measures available are listed below with the national values reported as of June 25, 2004 in parentheses. These measures are updated quarterly.

- Trucks involved in fatal crashes in the Crash file as a percent of the FARS file (100%).
- Percentage of crash reports uploaded within 90 days (64%).
- Percentage of inspections uploaded in 21 days (75%).
- Percent of interstate truck and bus crash records that can be matched with the Census file (87%)
- Percent of interstate truck and bus inspection records that can be matched with the Census file (96%).

Based on these measures, states are given a score of good, fair or poor. Thirteen states currently receive a poor rating. About 27 percent of trucks involved in fatal crashes are in these 13 states, based on 2002 FARS data. In addition, D.C., New Mexico, North Carolina and Vermont have been flagged in the Volpe tabulation (4) because less than half of their crashes for the previous year have been reported. These states are listed in Table 2.

**TABLE 2**  
**Data Quality Problem States Listed by Volpe(4) and UMTRI(5)**

<u>Volpe</u>	<u>UMTRI (Crash only)</u>
District of Columbia*	District of Columbia
Georgia	Alaska
Kentucky	Arizona
Louisiana	California
Minnesota	Florida
New Hampshire	Maine
New Jersey	Mississippi
New Mexico*	Nevada
New York	New Jersey
North Carolina*	New Mexico
Pennsylvania	North Carolina
Tennessee	Ohio
Vermont*	Oklahoma
	Texas
	Virginia

\*less than half of 2003 crashes received

The Center for National Truck and Bus Statistics at the University of Michigan Transportation Research Institute (UMTRI) has also been evaluating the MCMIS Crash file (5,6,7). UMTRI (5) describes trends in national totals from 1994–2000 comparing the MCMIS Crash file with UMTRI Trucks Involved in Fatal Accidents (TIFA), the National Highway Traffic Safety Administration (NHTSA) Fatality Analysis Reporting System (FARS) and the NHTSA General Estimates System (GES) for non-fatal crashes. UMTRI also addresses reporting by collision severity showing fatal, injury and tow-away national counts. UMTRI reports that the number of trucks involved in fatal crashes in the Crash file has improved to over 90 percent (93 percent in 2000) of the expected value. However, national totals from the Crash file for injury and tow-away crashes are about 80 percent and 50 percent respectively of the expected numbers. Overall reporting of truck crashes to MCMIS ranges from 62–67 percent of expected national counts.

State by state differences are also tabulated by UMTRI for 1998, 1999 and 2000 for trucks involved in fatal crashes. While the national totals from the Crash and TIFA files are quite close, individual states can differ by as much as 30 percent. In 2000, Texas was 27 percent low, but Tennessee was 35 percent high. The fatal count in the 2000 Crash file is only 7 percent below the TIFA file, but this figure arises from a combination of 31 states that were down collectively 10 percent and 13 states that were collectively up 3 percent (some states in Crash agreed exactly with FARS). UMTRI also identified problem states and these are also listed in Table 2 with the Volpe states. However, UMTRI only considered crash reporting over the period 1994–2000, while Volpe is evaluating several data quality measures using current data.

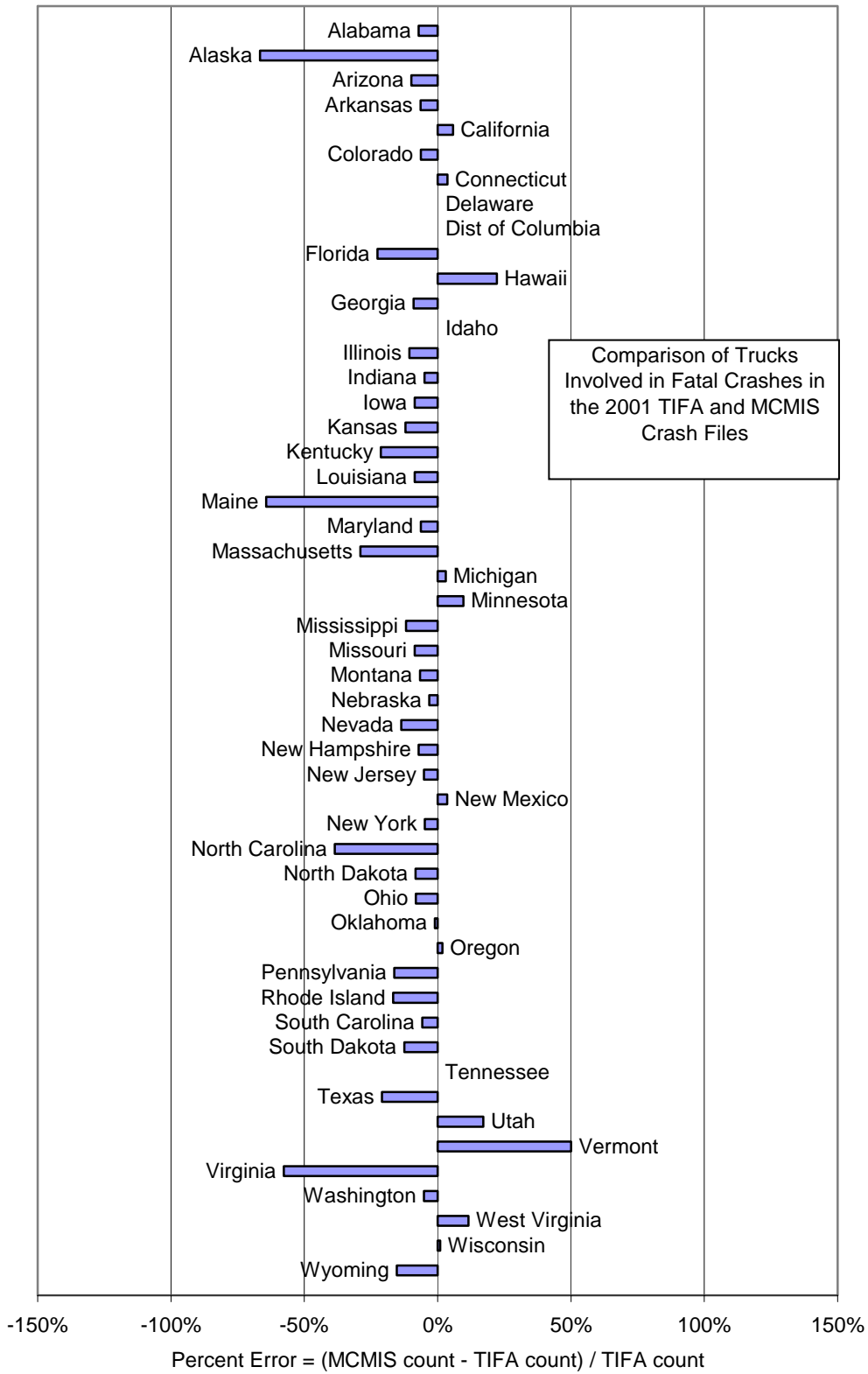
In a more detailed examination of Missouri data from 2001 (6), UMTRI found that about 20 percent of the records entered in the Crash file were duplicates. Corrected cases were added, but the original case was not deleted. After correcting for this, UMTRI determined that Missouri reported 77 percent of fatal crashes, 64 percent of injury crashes and 59 percent of tow-away crashes producing a total reporting to MCMIS for 2001 of 61 percent. While the reporting of fatal crashes to MCMIS is quite good, there is still substantial under-reporting of non-fatal crashes. Evidence of state-to-state differences in the reporting of fatal crashes suggests there are substantial differences between states in reporting non-fatal crashes to MCMIS.

Figure 1 compares counts of trucks in the 2001 TIFA<sup>1</sup> and Crash files. The difference in counts is shown as a percentage of the count in the TIFA file for each state. Several states are low by nearly 50 percent. However, the problem states are not always the same from year to year. For example, Tennessee was low by 40 percent in 1999, high by 35 percent in 2000 (perhaps the missing 1999 reports were counted in the wrong year) and exactly the same in 2001.

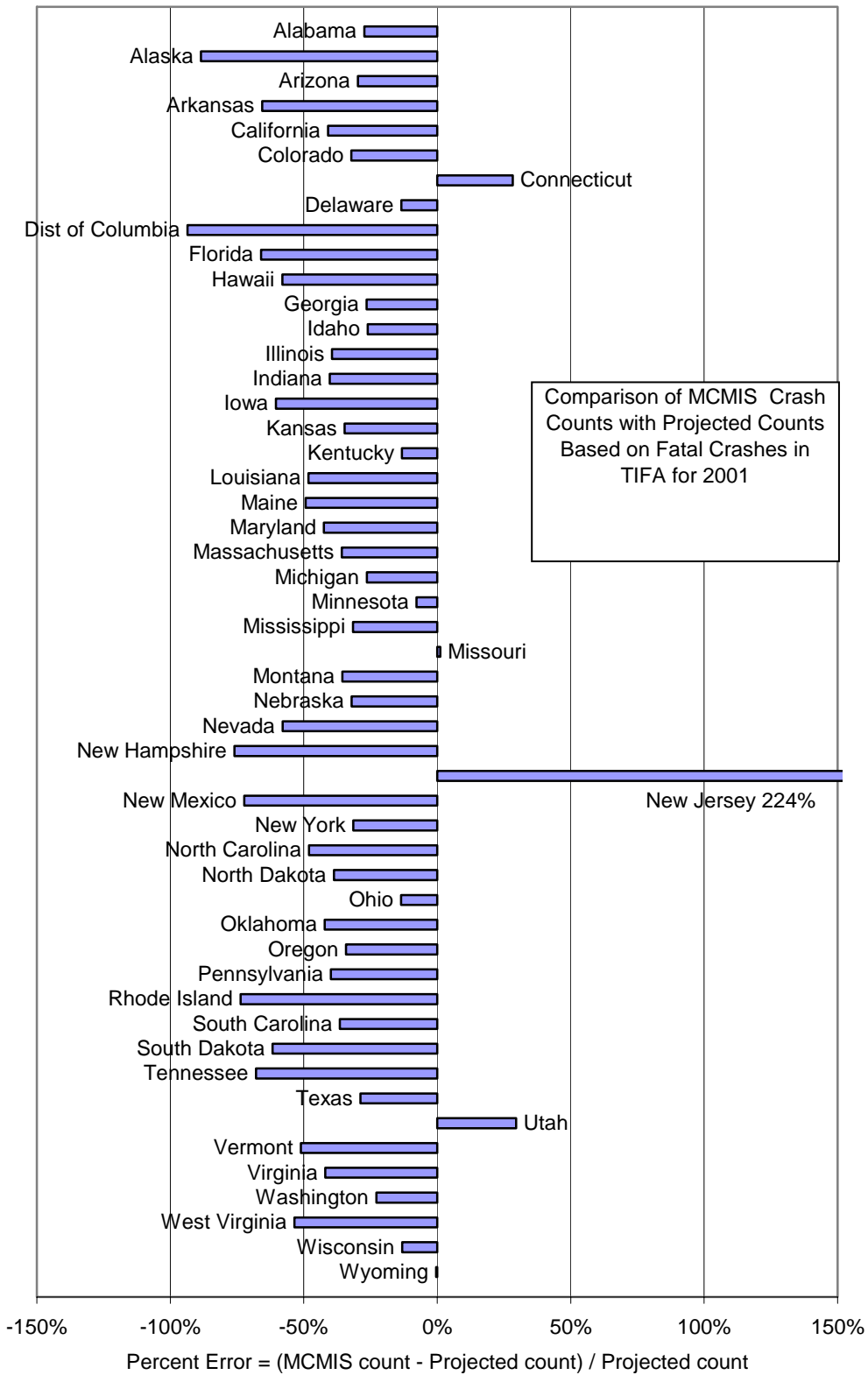
A state-by-state evaluation of missing records in the Crash file for non-fatal collisions was not addressed by Volpe or UMTRI. That is because the available national data on non-fatal crashes, GES, cannot be tabulated by state (GES is based on only 60 sampled areas in the U.S. and does not collect data in every state). An approximation of the total number of crashes expected in the MCMIS file is presented in Figure 2. The ratio of fatal crashes to tow-away crashes at the national level (about 31) can be calculated from the data in the UMTRI report (5). This ratio was used to inflate the fatal counts in TIFA for each state to estimate the total number of tow-away crashes that could be expected in the Crash file. The difference between the projected state counts and the actual count in MCMIS Crash file is shown as a percent of the projected count in Figure 2.

---

<sup>1</sup> A preliminary tabulation from the 2001 TIFA file was provided by UMTRI since the final version has not been released yet.



**Figure 1: Comparison of Trucks in Fatal Crashes in the 2001 TIFA and MCMIS Crash Files**

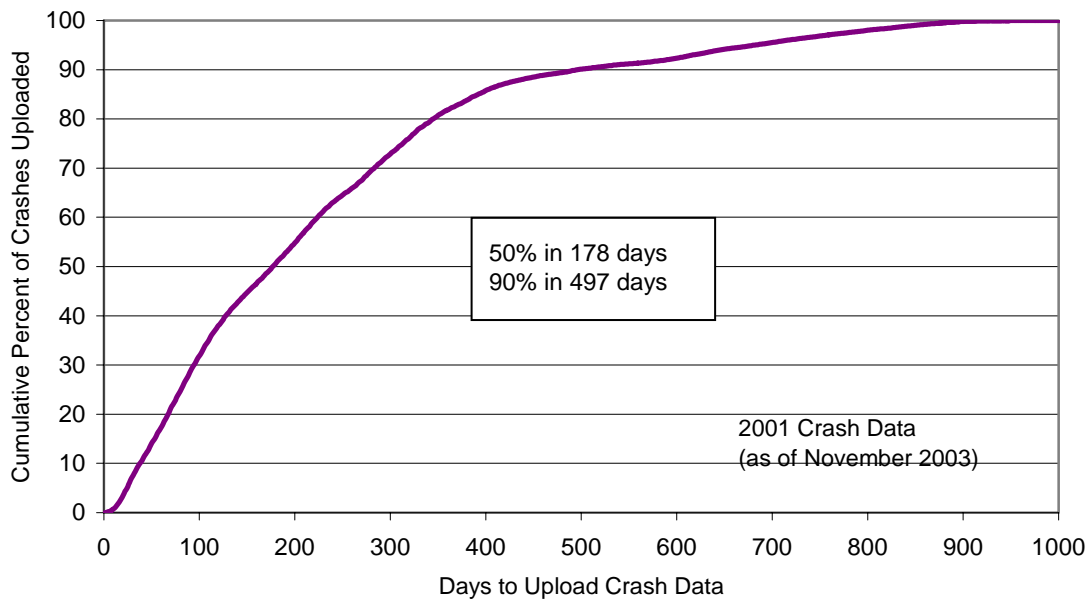


**Figure 2: Comparison of Projected and Actual Counts in the 2001 MCMIS Crash File**

Using this method, the projected national total is about 160,000 and the MCMIS Crash file total of 107,484 is about two-thirds of this. The MCMIS counts for many states are below the projected counts by 50 percent and more. An exception is New Jersey that is high by 224 percent. This may be another case of duplicate records. The projected counts for states with small counts of fatal crashes should be tempered by the fact that their fatality count can fluctuate substantially from year to year and the projection will suffer accordingly. Also, the ratio of fatal to tow-away crashes may vary from state to state. Contrary to the UMTRI report, Missouri looks very good in these figures. Fatal counts are now in good agreement. Duplicate records have apparently been removed and the total count agrees with this approximation. However, UMTRI's analysis of the Missouri police accident reports (6) produces a total that is more than 50 percent higher (7,356). These examples illustrate that it is not easy to determine what the expected number of reports in the Crash file should be. It also shows that the problems tend to shift from state-to-state and year-to-year as people work to resolve the known problems and new problems arise. Replicating the reliability of the FARS file in a database three times as large on a smaller budget is not an easy task. The next section addresses late data.

## **2.2 LATE DATA**

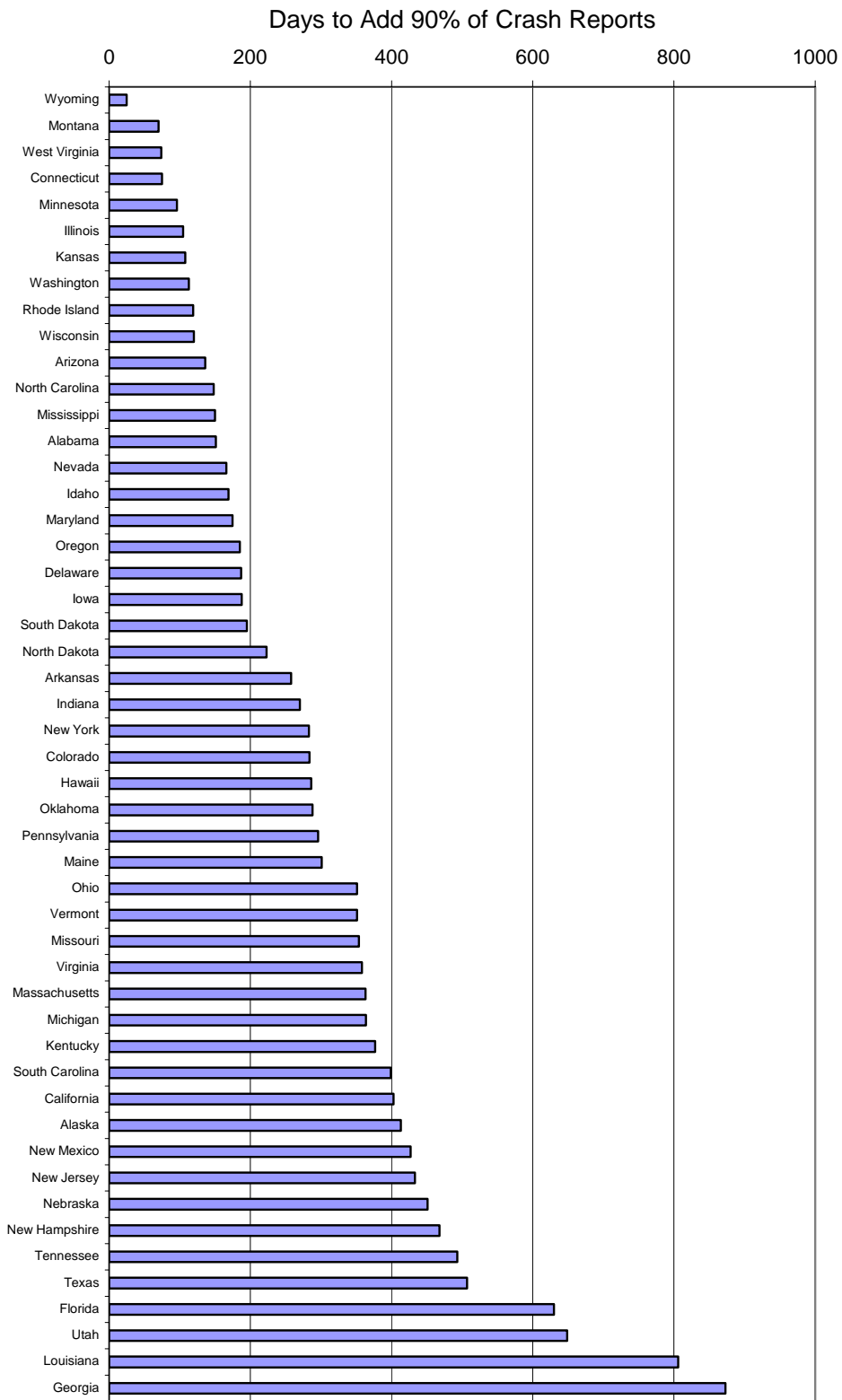
While the roadside inspection data are generally received within 21 days, timely receipt of the crash data is a problem that FMCSA has been working on for some time. The Volpe evaluation used data received as of March 2003 for the simulated ranking on March 2001. For this analysis, a November 2003 version of the MCMIS Crash file was used to look at the number of days it took to receive each crash occurring in calendar year 2001. Whereas the Volpe simulated ranking allows 24 months for late data to arrive, the file examined here includes 22 months after the end of the 2001 calendar year. The cumulative distribution of days to receive the 2001 crash data is shown in Figure 3. Fifty percent of the crashes were received in 178 days and 90 percent were received in 497 days. The average number of days to receive a 2001 crash was 228. This distribution can be used to calculate that about 25 percent of the crashes would be missing from a 30 month period ending at the current date, the window used by the SafeStat algorithm. This distribution can also be used to estimate that the post-selection period is missing about 17 percent of the crashes in that period because they were not received by MCMIS as of March 2003.



**Figure 3: Distribution of Time Required to Upload Crash Data**

The number of days to receive 90 percent of the 2001 crashes is shown by state in Figure 4. Nearly all states had 90 percent of their 2001 crashes in within 16 months , although D.C. had not submitted any reports as of November 2003. The variation from state-to-state is substantial and several of the slowest are large southern states such as Texas, Florida, Tennessee and Georgia. This illustrates a potential for geographic bias due to late data.

For 2002 crashes, FMCSA reports (8) that the average number of days to receive a 2002 crash decreased to 157 (based on data as of March 31, 2004). Volpe also reports current information on state data quality. Using data as of June 25, 2004, they report that 64 percent of crashes are now received in 90 days and that 75 percent of inspection records are received in 21 days (4).



**Figure 4: Days to Receive 90 Percent of 2001 Crash Data by State**

### 2.3 EFFECT OF LATE DATA ON CARRIER RANKINGS

The Volpe evaluation used data received by March 2003 for the simulated ranking as of March 2001. This effectively removes the impact of late data from the Volpe evaluation, which was focused on the effectiveness of the SafeStat algorithm in identifying high risk carriers. Volpe also provided ORNL with the results of the original ranking from March 2001. Comparison of these two rankings provides an example of the net impact of data for the 30-month ranking period (September 1998—March 2001) that was received up to 2 years late (March 2003). There is one other difference in the two rankings. The simulated ranking used the current SafeStat algorithm, Version 8.5, while Version 8.2 was in use in March 2001. Differences between the two versions are summarized in Table 3. These differences could be responsible for some changes in rankings between the two runs, but most changes are presumed to be due to the substantial amount of added data.

**TABLE 3**  
**Changes in SafeStat versions used for the Original and Simulated March 2001 runs**

March 2001	Simulated March 2001
Version 8.2 (as of March 2001)	Version 8.5 (as of January 2003)
Enforcement History Indicator (EHI) is limited to only using data from enforcement cases initiated by compliance reviews (CR) or terminal audits.	EHI use all closed enforcement cases, including those not initiated from CRs or terminal audits. (change made March 2002)
Power unit (PU) figures used in the Accident Involvement Indicator (AII) calculation.	PU used in AII is an <b>average</b> of the carrier's PU totals at the end of the three time periods used for time-weighting in the AII calculation (0-6, 6-18, and 18-30 months). (change made March 2002)
Trip-leased PU not included.	Trip-lease PU numbers are added to the owned and term-leased PUs to determine the total number of PUs of a carrier. (change made March 2002)
SafeStat use only the violations of acute and critical regulations that were used as part of the safety rating as defined in Part 385 Appendix B of the FMCSR.	Hazardous Material Review Indicator (HMRI) was expanded to include non-ratable violations of acute and critical regulations. (change made September 2001)
	Indicators in the Accident SEA (both the AII and Recordable Accident Indicator) based on only one crash will not be calculated. (change made January 2003)

Note: Volpe stated that, in its simulation run conducted for the effectiveness study that the average power unit calculation used in the Accident Involvement Indicator (AII) in version 8.5 was not employed (see Appendix A). This removes one of the differences between versions.

The SafeStat scores are translated into letter categories, as shown in Table 4. Only category A and B carriers are identified as at-risk. Category C carriers are classed as “other identified.” All other carriers are characterized as “not identified, including the vast majority with insufficient data to receive any SafeStat score. The data provided by Volpe for this review provide a unique opportunity to assess the impact of late data. Table 5 shows the number of carriers in each of the SafeStat letter categories in the March 2001 (using the data available as of March 2001) ranking versus the Safestat category when all data received as of March 2003 are used to rank carriers as of March 2001.

**TABLE 4**  
**Definition of SafeStat Categories**  
**(includes carriers with SEA values of 75 or higher):**

<b>Carriers Identified</b>		
<b>Category</b>	<b>SafeStat Score</b>	
<b>A</b>	350 to 550	All 4 SEAs 3 SEAs that result in a weighted score of 350+
<b>B</b>	225 to less than 350	3 SEAs that result in a weighted score < 350 2 SEAs that result in a weighted score of 225+
<b>C</b>	150 to less than 225	2 SEAs that result in a weighted score <225
<b>Carriers Not Identified, but with a Single SEA&gt;75</b>		
	<b>SEA Value</b>	<b>Specific SEA</b>
<b>D</b>	75 -100	Accident
<b>E</b>	75 – 100	Driver
<b>F</b>	75 – 100	Vehicle
<b>G</b>	75 – 100	Safety Management

**TABLE 5**  
**Number of Carriers by SafeStat Category:**  
**Original versus Simulated March 2001 Runs**

<b>SafeStat Categories</b> <b>March 2001 Run</b>	<b>SafeStat Categories in the Simulated March 2001 Run</b>							
	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>G</b>	<b>All</b>
<b>A</b>	300	50	19	1	4	4	1	379
<b>B</b>	214	2,753	219	51	145	125	26	3,533
<b>C</b>	132	229	2,084	0	183	165	67	2,860
<b>D</b>	2	60	0	1,448	14	20	8	1,552
<b>E</b>	14	401	213	15	8,127	22	3	8,795
<b>F</b>	7	285	204	24	24	14,606	8	15,158
<b>G</b>	4	63	59	1	2	1	1,639	1,769
<b>H</b>	4	88	39	1,054	918	1525	118	3,746
<b>I</b>	1	7	1	176	53	153	23	414
<b>All</b>	678	3,936	2,838	2,770	9,470	16,621	1,893	38,206

The table covers all 38,206 carriers that received a SafeStat category using the March 2003 data. Categories H and I were added by Volpe to the March 2001 data for carriers with no SEA values or an SEA value less than 75. The diagonal entries of the table show the numbers of carriers with the same category in both runs. The additional data did not change their ranking. The following observations can be made from Table 5.

- SafeStat categories for 30,957 carriers stayed the same between the two runs (81 percent of the total 38,206).
- SafeStat categories for 1,135 carriers changed from a higher category in the 2001 run to a lower category in the 2003 run (3 percent of 38,206). Among them, 595 carriers were dropped from the “SafeStat at risk” group (i.e., categories A and B), or 15 percent of the at-risk carriers in the March 2001 run.
- SafeStat categories for 1,954 carriers changed from a lower category to a higher one (5 percent of 38,206). The number of A-B, at risk, carriers increased by 1,297 (33 percent) by the movement of carriers that were not at-risk in the 2001 run.
- The net change in the number of at-risk carriers is an increase of 702, 18 percent more than the 2001 number.
- 4,160 carriers that previously did not receive SafeStat values (which were coded as H or I) were assigned to SafeStat categories A-G by the 2003 run (11 percent of 38,206).
- Among those 4,160 carriers with newly assigned SafeStat categories, 100 carriers received Category A or B; 40 received Category C; and 4,020 received single SafeStat Categories of D - G.

The crash data are generally received later than other data and crash data have the greatest impact on the carrier rankings. Accordingly, Tables 6 and 7 show the weighted number of crashes for the 30-month period, September 1998 to March 2001, by the SafeStat category of the carrier. The crash data used for the March 2001 run are shown in Table 6 and for the March 2003 run in Table 7. Finally, the difference in the number of crashes in each cell of the table is shown in last Table, 8.

**TABLE 6**  
**Total Weighted Number of Crashes from March 2001 Run**

SafeStat Categories March 2001 Run	SafeStat Categories in the Simulated March 2001 Run							
	A	B	C	D	E	F	G	All
A	3,904	625	238	52	88	48	18	4,973
B	1,863	15,594	1,023	1,008	1,692	1,204	194	22,578
C	1,273	777	5,951	0	346	415	211	8,973
D	40	800	0	26,843	247	374	51	28,355
E	87	4,173	357	923	19,449	24	21	25,034
F	39	1,656	433	120	36	19,276	68	21,628
G	46	2,675	174	9	8	0	4,403	7,315
H	13	456	41	12,741	1,620	1,708	558	17,137
I	0	2	0	394	7	24	11	438
All	7,265	26,758	8,217	42,090	23,493	23,073	5,535	136,431

**TABLE 7**  
**Total Weighted Number of Crashes from the Simulated March 2001 Run**

SafeStat Categories March 2001 Run	SafeStat Categories in the Simulated March 2001 Run							
	A	B	C	D	E	F	G	All
A	4,693	649	251	79	99	44	1	5,816
B	2,802	18,746	1,235	1,167	1,850	1,106	152	27,058
C	2,174	1,076	7,654	0	473	645	278	12,300
D	44	849	0	31,766	258	365	58	33,340
E	169	6,815	559	1,389	26,390	46	29	35,397
F	63	3,228	628	258	52	25,247	112	29,588
G	67	4,023	215	23	8	0	5,638	9,974
H	35	1,055	85	23,192	2,230	2,421	745	29,763
I	6	31	0	1,357	11	73	11	1,489
All	10,053	36,472	10,627	59,231	31,371	29,947	7,024	184,725

- Total weighted number of crashes increased from 136,431 in 2001 run to 184,725 in 2003 run. The “late data” added 48,294 crashes which is a 35 percent increase over the 2001-run level, or 26 percent of the March 2003 total.
- Specifically, the total weighted number of crashes increased from 95,420 in the 2001 run for those carriers that stayed in the same SafeStat categories (yellow zone) to 120,134 crashes in 2003. The “late” data, in this case, added about 26 percent of the 2001 total (24,714 crashes). As expected, the largest increase (percentage) is for those originally in the **H** and **I** categories. The total number of crashes increased from 17,575 to 31,252 between the 2 runs, which is about 78 percent over the 2001 level.
- The “red zone”, carriers that moved up in categories, went from 15,489 crashes in 2001 run to 24,442 in 2003 run. That’s about 58 percent increase from the 2001

numbers. The “green zone”, those that drop down in categories, went from 7,947 in 2001 run to 8,897 in 2003-run. This is the least changed group, but still added 12 percent over the 2001-run numbers.

**TABLE 8**  
**Difference in the Number of Crashes**

SafeStat Categories March 2001 Run	SafeStat Categories in 2003 Run							
	A	B	C	D	E	F	G	All
A	789	24	13	27	11	-4	-17	843
B	939	3,152	212	159	158	-98	-42	4,480
C	901	299	1,703	0	127	230	67	3,327
D	4	49	0	4,923	11	-9	7	4,985
E	82	2,642	202	466	6,941	22	8	10,363
F	24	1,572	195	138	16	5,971	44	7,960
G	21	1,348	41	14	0	0	1,235	2,659
H	22	599	44	10,451	610	713	187	12,626
I	6	29	0	963	4	49	0	1,051
All	2,788	9,714	2,410	17,141	7,878	6,874	1,489	48,294

For the most part, the number of crashes increased in each cell. However, the exclusion of single crashes from the accident SEA in the Version 8.5 of the SafeStat algorithm may have had a significant impact on these tables. The few negative cells in Table 8 may be due to this change in the algorithm.

Overall, the impact of late data on the ranking is significant. It is generally true that most of the changes with the addition of late data move additional carriers into the at-risk group, resulting in a 33 percent increase in at-risk carriers. However, 15 percent of the at-risk carriers in the March 3001 run moved out of the at-risk group with the addition of the late data. Missing data resulted in these carriers being mistakenly classified as at-risk in the original run. It is also possible that changes in the algorithm (SafeStat version) used are responsible for the movement of some of these carriers, if, for example, they had only one crash. The net change in the number of at-risk carriers is an increase of 18 percent over the March 2001 number.

#### 2.4 THE IMPACT OF MISSING OR LATE DATA

The original Volpe SafeStat evaluation (3) argues that missing crash data do not bias the rankings. The information reviewed here does not support that assumption. Of course, the problem with missing data is that it is missing and one doesn't know much about it. However, good national data are available on fatal truck crashes. These data reveal substantial state-to-state variation even though the total count of fatal reports in the Crash file is reasonably complete. Bias is also shown in the missing data distribution with collision severity. Only about half of tow-away crashes are submitted. A projected distribution also suggests substantial state-to-state variation in the reporting of non-fatal crashes. Examination

of underreporting from year-to-year shows a changing pattern from state-to-state. A substantial group of southern states are among the slowest to submit crash data. It is likely that there are state-to-state patterns in enforcement and traffic violation data as well. If all carriers operated uniformly in every state, then these variations might not affect the rankings. However, the majority of carriers are regional and few operate in a uniform fashion across the entire U.S. For many carriers, it is likely that missing data may bias their SafeStat ranking.

The assessment of the impact of late data only—missing data are still missing—shows that the ranking for the vast majority of carriers, 81 percent, did not change. However, there were significant changes in the important at-risk group with 15 percent moving out of that group with the addition of late data. In this case, late data were about 26 percent of the total number of cases. Overall underreporting to the MCMIS Crash file is estimated at an additional one-third of the total number of cases, or half of the known cases. It is likely that these additional missing records also bias the rankings, depending on whether a carrier operates in areas with good reporting or not.

### 3 STATISTICAL METHODS

This section addresses the SafeStat algorithm. In order to evaluate something, it is helpful to have a comparison. The SafeStat algorithm is the product of expert judgment. This review makes the assumption that the most likely alternative is to replace expert judgment with statistical methods. SafeStat is a formula that reflects assumptions about the relationships of various measures to crash risk. The same source data could support statistical models that can also relate the source data to crash risk. Thus, the focus of Section 3 is to make a controlled comparison of the ability of SafeStat and several standard statistical methods to predict carrier risk. A standard evaluation tool for statistical methods is to compare the outcome as determined by the method with the actual outcome for each carrier and this approach provides an alternative way to evaluate both SafeStat and the statistical methods.

The material in this section demonstrates (1) that straightforward statistical methods can substantially improve the SafeStat carrier classification, (2) that alternatives to Volpe's method for evaluating effectiveness are more informative, and (3) that although classifying carriers on the basis of SEAs and similar indicators of safety is effective at identifying high-risk carriers, the statistical limitations of the classifications should be better understood and addressed.

**Background.** The Volpe SafeStat method scores motor carriers on the basis of four Safety Evaluation Areas (SEAs) relating to crashes, drivers, vehicles, and safety management. The four SEAs are combined into an overall safety risk score, which is used to classify carriers into safety risk categories. The primary objective in the classification is to target high-risk carriers for inspections and compliance reviews, both of which could potentially prevent crashes.

An updated effectiveness study (3) was performed by Volpe to evaluate the SafeStat method. In the study, carriers are first classified by the SafeStat method on the basis of carrier data from the 30 months prior to March 2001, and this classification is then compared to actual post-selection carrier performance data for the 18 months following March 2001. Three safety risk categories are of particular importance in the study: "At-Risk (Category A&B)," "Other Identified (Category C)" and "Not Identified." Usable and ample post-selection data was available for 118,757 carriers, of which, according to the prior-data SafeStat classification, 3,595 were in the At-Risk Category A&B, 2,357 were in the Other Identified Category C and 112,805 were Not Identified. The combination of Categories A, B and C are referred to as "All Identified."

Volpe provided ORNL with the 30-month "pre-selection" and 18-month "post-selection" data and results from the SafeStat algorithm. Using these data, Volpe's Post-Selection Crash Rates were reproduced and are shown in Table 9. From the "percent higher than not identified carriers" figures (last column in Table 9), Volpe concluded that the SafeStat method is effective. The next step is to see if other alternatives to the SafeStat classification and the Volpe crash-rate evaluation method used for Table 9 can illustrate shortcomings in the current method.

**Table 9**  
**SafeStat Post-Selection Crash Rates**  
**with Power Units and Weighted Crashes**

<b>Carrier Group</b>	<b>Carriers</b>	<b>Weighted Crashes</b>	<b>Power Units</b>	<b>Weighted Crashes per 1000 PUs</b>	<b>Percent Higher Than Not Identified</b>
All	118,757	58,607.8	2,293,655	25.55	4.0
All SafeStat Identified	5,952	5,363.5	126,152	42.52	73.1
SafeStat At-Risk (A&B)	3,595	3,805.0	73,207	51.98	111.6
Other SafeStat Identified (C)	2,357	1,558.5	52,945	29.44	19.8
SafeStat Not Identified	112,805	53,244.3	2,167,503	24.56	0.0

### 3.1 REGRESSION MODELS

The SafeStat scores are based on expert judgment and make sense intuitively, but they do not take advantage of standard statistical methodology. For example, the four SEAs are combined into an overall SafeStat score using the subjectively determined weights 2, 1.5, 1, and 1 for safety, driver, vehicle, and safety management SEAs respectively. Therefore, one of the first steps in this evaluation was to see if statistically derived weights might lead to a better classification. This was accomplished by performing a Poisson regression of the SafeStat pre-selection weighted total crashes onto the four SEAs and the number of power units (PUs).

Although the dependent variable in the Poisson regression is total crashes, the regression is based on crash rates (weighted crashes per PU) as well, because the number of PUs is a term in the regression model.<sup>2</sup> In order to provide a direct comparison to the SafeStat ranking, the fitted regression equation was used to rank the same 118,757 carriers used in the Volpe study (3), according to their regression-predicted crash rates (weighted crashes per PU). The 3,595 carriers with the highest regression-predicted risk compose an alternative to the SafeStat at-risk carriers, and the 2,357 carriers with the next highest regression-predicted risk compose an alternative to the SafeStat Other Identified carriers. The remaining 112,805 carriers compose an alternative to the SafeStat Not Identified carriers. The actual post-selection data were then used to compute crash rates for the regression-based categories. Table 10, computed for the regression-based classification, is the direct analog of Table 9 for the SafeStat classification.

<sup>2</sup>In the Poisson regression model, the  $\mu$ , the mean total weighted crashes is expressed as  $\mu = a \times (\#PUs)^b \times \exp [c \times (\text{Accident SEA}) + d \times (\text{Driver SEA}) + e \times (\text{Vehicle SEA}) + f \times (\text{Safety Management SEA})]$ , where a-f are parameters to be estimated. Since both sides of the model equation can be divided by the number of PUs, the regression is also based on the weighted crash rate.

From the last columns in Tables 9 and 10, it can be seen that the Poisson model classification is considerably more effective than the SafeStat classification, according to the Volpe “percent higher than not-identified” metric. One reason the Poisson procedure is better is its relative treatment of the four SEAs. Though all four SEAs are statistically significant in the regression, the fitted regression coefficients for the accident, driver, vehicle, and safety management SEAs are in the relative proportions of 57: 3.2: 1.4: 1. Thus the Poisson regression attributes much more weight to the accident SEA in the classification, than the SafeStat formula.

**Table 10**  
**Post-Selection Crash Rates**  
**for Poisson Model Carrier Groups**

<b>Carrier Groups Based on Poisson Model</b>	<b>Carriers</b>	<b>Weighted Crashes</b>	<b>Power Units</b>	<b>Weighted Crashes per 1000 PUs</b>	<b>Percent Higher Than Not Identified</b>
All	118,757	58,607.8	2,293,655	25.55	8.5
All Identified	5,952	7,215.3	111,584	64.66	174.5
At-Risk (A&B)	3,595	2,567.8	36,995	69.41	194.7
Other Identified (C)	2,357	4,647.5	74,589	62.31	164.6
Not Identified	112,805	51,392.5	2,182,071	23.55	0.0

Another reason the Poisson procedure is better is the inclusion of the number of power units as a term in the regression model. According to the fitted model parameters, weighted crash rates are slightly smaller for larger carriers,<sup>3</sup> perhaps because economy of scale makes safety more affordable for them. The SafeStat method does not formally account for an effect of carrier size (number of power units) on crash rates.

**Another Approach.** Larger carriers represent greater potential for reducing crash frequency. The average fleet size for A&B carriers identified by the Poisson model is only about 10, as compared to an average fleet size of about 20 for A&B carriers identified by SafeStat. Small carriers are statistically more variable and thus have a tendency to have both higher and lower crash rates merely because of random variation. Another statistical method was tried to address this issue.

A second regression was performed with a lower confidence bound (LCB) for total weighted crashes instead of total weighted crashes as the regression dependent variable. Because a high crash rate may be only a statistical anomaly, carriers identified as high-risk should be

---

<sup>3</sup>The estimate of b (see footnote 2) is less than 1.

those with high crash rate lower confidence bounds. In the sense that large-carrier crash rates are statistically more stable, LCBs for rates tend to be higher for larger carriers. Approximate lower confidence bounds for total weighted crashes were computed by approximating the total weighted crash distribution as Poisson.<sup>4</sup>

Table 11 presents the results computed from the LCB-based regression model. It is clear from the values in the Weighted Crashes and Power Units columns in Tables 11 that the LCB-based regression tends to select carriers with more power units and more weighted crashes than either the rate-based regression (Table 10) or the SafeStat method (Table 9). Like the rate-based regression, the LCB-based regression also outperforms the SafeStat classification, according to the “percent higher than not-identified” statistics.

**Table 11**  
**Post-Selection Crash Rates**  
**for LCB Model Carrier Groups**

<b>Carrier Groups Based on LCB Model</b>	<b>Carriers</b>	<b>Weighted Crashes</b>	<b>Power Units</b>	<b>Weighted Crashes per 1000 PUs</b>	<b>Percent Higher Than Not Identified</b>
All	118,757	58,607.8	2,293,655	25.55	12.5
All Identified	5,952	10,415.0	171,980	60.56	166.6
At-Risk (A&B)	3,595	4,735.5	69,530	68.11	199.8
Other Identified (C)	2,357	5,679.5	102,450	55.44	144.1
Not Identified	112,805	48,192.8	2,121,675	22.71	0.0

**Alternative Evaluation Method.** If the goal of the SafeStat classification system is to identify high-risk carriers and to target those carriers for inspections or compliance reviews, then an objective should be to minimize misclassification of individual carriers. Incorrectly identifying carriers as high-risk may cause unnecessary concern for carriers and can waste auditing and inspection resources that could be assigned to true high-risk carriers. On the other hand, unsafe carriers who are passed-over by the algorithm may have future crashes that could be prevented by inspections or reviews. The statistics in Tables 9-11 do not show whether individual carriers identified as at-risk actually were high-risk in the post-selection period. The “percent higher than not identified” are aggregate statistics for the entire At-Risk, Other Identified, or Not Identified groups. For example, it is possible that all of the crashes for the At Risk group occur for just a few carriers and that most of the carriers classified in the group do not belong there.

---

<sup>4</sup>Computed with the inverse gamma function as exact Poisson confidence bounds.

Carriers who operated during the post-selection period can be classified according to their post-selection weighted crash rates or crash-rate LCBs. Using the post-selection crash-rates, the 118,757 carriers were ranked and classified into At-Risk, Other Identified, and Not Identified actual-outcome categories. The 5,952 carriers with the highest post-selection crash rates were classified into the All Identified category, and the remaining 112,805 were classified as Not Identified. The 3,595 carriers with the highest post-selection crash rates were classified into the At-Risk category. For comparisons with LCB-based method, LCBs for both post-selection and pre-selection rates were also used, in place of raw rates. That is, the 5,952 carriers with the highest actual post-selection LCBs were selected for the LCB-based All Identified actual-outcome category, 3,595 were selected for the At-Risk actual-outcome category, and 112,805 were classified as Not Identified.

The SafeStat and regression classifications were then compared with the actual post-selection outcomes. The comparisons were made in terms of 2x2 classification tables. Table 12 is a generic 2x2 classification table. The table entries are:

- A the number of carriers who were not identified as high-risk by the method and were not high-risk in the post-selection period;
- B the number of carriers who were not identified as high-risk but were high-risk in the post-selection period;
- C the number of carriers who were selected as high-risk but were not high-risk in the post-selection period; and
- D the number of carriers who were selected as high-risk and were also high-risk in the post-selection period.

Cells A and D both represent outcomes that are consistent with the classification made by the method. Cells B and C are both outcomes that are inconsistent with the initial classification. One classification method is better than another when the proportion of correct classifications is larger.

**Table 12**  
**A Generic Classification Table**

Selected by Method?	Actual Outcome: Post-Selection Crash Rate is High?	
	No	Yes
No	A (Correct classification of low risk carriers)	B (Incorrect classification)
Yes	C (Incorrect classification)	D (Correct classifications of high-risk carriers)

Tables 13 and 14 contain 2x2 classification tables comparing the SafeStat and regression classification methods to the actual post-selection outcome. Before considering these tables, however, it is useful to observe a few mathematical properties about them. The SafeStat and regression methods select either 3,595 or 5,952 carriers for the At-Risk and All Identified

groups. Therefore, in each table, C+D is either 3,595 or 5,952. Further, A+B, the number not selected, is either 115,162 (118,757 – 3,595) or 112,805 (118,757 – 5,952). Because we have also taken either 3,595 or 5,952 as the number of actual-outcome carriers that should be selected, B+D is also either 3,595 or 5,952. It follows that B+D = C+D, and thus B = C.

It also follows from these conditions that any one of A, B, C, or D uniquely determines all the others. All of the table properties can therefore be inferred from any one of A, B, C, or D, or from single statistics computed from A, B, C, and D. One such statistic, which is called the *lift*, is the ratio P/Q of two proportions P and Q, where P is  $D/(D+C)$ , the proportion of should-have-been-selected carriers among those that were selected, and Q is  $(B+D)/(A+B+C+D)$ , the proportion of should-have-been-selected carriers among all carriers. The lift represents the increase in efficiency, relative to sampling at random, in finding carriers that should have been selected in the At-Risk or All Identified categories. Under the conditions on the tables that hold here, the more lift, the better the classification method.

Tables 13 and 14 show that for either Poisson rate-based or LCB-based classifications, the regression approach gives more lift than the SafeStat method. Alternatively, the approaches can also be compared in terms of any one of A, B, C, or D. Consider, for example, D, the number of selected carriers that were also high-risk in the post-selection period. In both cases, D is higher for the regression classification. For example, Table 13 shows that for the At-Risk classification, the SafeStat method correctly selected 219 high-risk carriers, while the Poisson method correctly selected 265 (with both methods selecting a total of 3,595). Thus the Poisson method would have identified 46 (21 percent) more unsafe carriers than the SafeStat method.

**Statistical Limitations.** Although the lift values in Tables 13 and 14 are all around 2 – 2.5, the actual correct identification rates for high-risk carriers are all in the 6-12 percent range. Between 88 and 94 percent of selected at-risk carriers were not high-risk in the following 18 months! The selection methods double the identification rate of high-risk carriers, which means they are much more efficient than random sampling at screening for unsafe carriers, but still only about 1 selected carrier in 10 actually is high-risk in the post-selection period. This is a very substantial limitation. The implication is that most of the carriers identified as at-risk by SafeStat were selected due to random variations in the source data.

By providing statistical lift, the SafeStat scores are a useful screening tool. Classification scores should only be quoted with confidence limits that properly qualify their uncertainty, and classification methods should always be interpreted in the context of their statistical limitations. The SafeStat methodology does not currently incorporate these functions. Another advantage of statistical methods is that the uncertainty, or variability, of the result is also estimated. This is addressed in the next section.

**Table 13**  
**SafeStat and Regression Classifications**

<b>Carrier Classification for SafeStat At-Risk Category</b> <b>Number Selected = 3,595, Lift = 2.01</b>			
Post-Selection Crash Rate is High?			
Carrier Selected by Method?	No	Yes	Percent Selected
No	111,786	3,376	2.93
Yes	3,376	219	6.09
<b>Carrier Classification by Poisson At-Risk Category</b> <b>Number Selected = 3,595, Lift = 2.44</b>			
Post-Selection Crash Rate is High?			
Carrier Selected by Method?	No	Yes	Percent Selected
No	111,832	3,330	2.89
Yes	3,330	265	7.37
<b>Carrier Classification by SafeStat All Identified Category</b> <b>Number Selected = 5,952, Lift = 1.97</b>			
Post-Selection Crash Rate is High?			
Carrier Selected by Method?	No	Yes	Percent Selected
No	107,441	5,364	4.76
Yes	5,364	588	9.88
<b>Carrier Classification by Poisson All Identified Category</b> <b>Number Selected = 5,952, Lift = 2.15</b>			
Post-Selection Crash Rate is High?			
Carrier Selected by Method?	No	Yes	Percent Selected
No	107,495	5,310	4.71
Yes	5,310	642	10.79

**Table 14**  
**SafeStat and LCB-Based Classifications**

<b>Carrier Classification by SafeStat At-Risk Category</b> <b>Number Selected = 3,595, Lift = 2.24</b>			
Post-Selection Crash Rate is High?			
Carrier Selected by Method?	No	Yes	Percent Selected
No	111,811	3,351	2.91
Yes	3,351	244	6.79
<b>Carrier Classification by LCB At-Risk Category</b> <b>Number Selected = 3,595, Lift = 2.68</b>			
Post-Selection Crash Rate is High?			
Carrier Selected by Method?	No	Yes	Percent Selected
No	111,859	3,303	2.87
Yes	3,303	292	8.12
<b>Carrier Classification by SafeStat All Identified Category</b> <b>Number Selected = 5,952, Lift = 2.05</b>			
Post-Selection Crash Rate is High?			
Carrier Selected by Method?	No	Yes	Percent Selected
No	107,464	5,341	4.73
Yes	5,341	611	10.27
<b>Carrier Classification by LCB All Identified Category</b> <b>Number Selected = 5,952, Lift = 2.28</b>			
Post-Selection Crash Rate is High?			
Carrier Selected by Method?	No	Yes	Percent Selected
No	107,532	5,273	4.67
Yes	5,273	679	11.41

### 3.2 CONFIDENCE INTERVALS FOR CRASH RISK SCORES AND RANKS

Motor carrier risk evaluation is similar to evaluating scholastic aptitude, and it is useful to consider similarities between the two kinds of evaluations. Scholastic aptitude tests (SAT) and SafeStat scores both come in several varieties (verbal and math SAT scores, accident or vehicle SEA scores), and both kinds of scores can be a very sensitive issue to score recipients. SAT scores or percentile ranks<sup>5</sup> are usually reported with confidence intervals (CIs), which serve two purposes: they keep the SAT scores or percentile ranks more credible by qualifying them with honest uncertainty bounds, and they soften the interpretation of the scores by score recipients. CIs in carrier risk reporting would be useful for the same reasons.

However, differences between SAT and SafeStat scores affect how CIs for them should be computed. For example, SAT scores are for individuals, whereas SafeStat scores are for carriers, which can comprise many individual power units (and drivers). Therefore normalization by number of power units (or drivers) is necessary in interpreting SafeStat scores and corresponding CIs. The CIs for the carrier risk scores discussed here will be ranges for individual power units.

A more fundamental difference between SAT and SafeStat scores has to do with test standards. For SAT scores, a standard “aptitude” is never really defined other than in terms of performance on the SATs themselves. Aptitude could be defined in terms of criteria such as performance in college, for example, but that is not done in determining SAT CIs. A “true” score is the average score a student would achieve in taking many substantially similar tests at about the same time.<sup>6</sup> Therefore SAT CIs are based on variation among test results for students given multiple, similar tests. CIs for individual (e.g., verbal) and overall scores (e.g., math plus verbal) are both computed this way. For these CIs, the only standard is performance on similar tests. The statistical error that is accounted for is the test-to-test differences.

In carrier risk prediction, on the other hand, there is a well defined standard. Actual (post-selection) carrier risks are used to judge the accuracy of the SafeStat scores, and actual carrier risks can also be used in to evaluate CIs. Thus “Estimate – Actual” prediction error has to be considered in evaluating CIs.

**Confidence Intervals for Risk Scores.** The prediction error of a carrier risk score can be decomposed into two components:

$$\text{Estimate} - \text{Actual} = (\text{Estimate} - \text{Expected}) + (\text{Expected} - \text{Actual})$$

Where:

---

<sup>5</sup>The *percentile rank* of a score having rank R among N scores is  $100 \times R/N$ .

<sup>6</sup> See for example, <http://www.collegeboard.com/counselors/hs/sat/aboutI/satfaq.html#howaccurate>.

the estimated value is the value predicted by the fitted model,  
the actual value is the observed value, and  
the expected value is the true value without random error.

Even if the expected carrier risk is unknown, we can still consider this decomposition. The first component (Estimate - Expected) of the total prediction error varies because of limitations in relating SEAs to long-term crash rate averages. The limitations are due to (i) statistical error in estimating the assumed model (e.g., in estimating the regression parameters), and (ii) error in the assumed model itself (i.e., model lack of fit). In CIs calculated for regression estimates, model lack of fit is generally not separated from overall statistical error.

The second component (Expected - Actual) of the overall prediction error varies because of short-term differences between crash rates and their long-term averages. If, for a given carrier, the period of observation is short-enough and the number of power units is small enough, then short-term variability will make the second error component the dominant error for that carrier.

Depending on the statistical model, contributions to the overall error can be estimated separately for the two components of the overall prediction error. For example, in least squares regression with normal regression errors, the statistical distribution of the (Estimate - Expected) component can be derived exactly. The statistical distribution of the (Expected - Actual) component for a new observation is also normal and statistically independent of the first term, and confidence limits can thus be derived for the overall prediction error (Estimate - Actual).

The mathematics for CIs works out exactly in least squares regression with normal errors. CIs can also be derived in other settings, though the derivations usually depend on large-sample or other approximations, and there are usually various ways to apply such approximations. Prediction confidence bounds can be computed from the regression models and the assumption that the underlying crash frequencies have Poisson distributions (or other, such as the negative binomial). The assumption of a Poisson distribution is itself an approximation here, particularly as applied to the total weighted crash variable (tctwa), which is a weighted sum of different crash counts and therefore not strictly Poisson.<sup>7</sup> Approximate CIs for Poisson regressions are derived in Appendix B ***The confidence bounds considered here are used as illustrations only. It is not a purpose of this review to derive confidence bounds that are somehow optimal.***

Table 15 shows 95 percent lower and upper prediction bounds for a few carriers. Notice, for example, carrier 261450, whose had 1 power unit and a predicted rate of .43 weighted crashes per power unit during the pre-selection period. The 95 percent prediction LCB

---

<sup>7</sup>Though beyond the scope of this review, crash rates could also be predicted and approximate CIs could also be computed using nonparametric analogs of least squares regression.

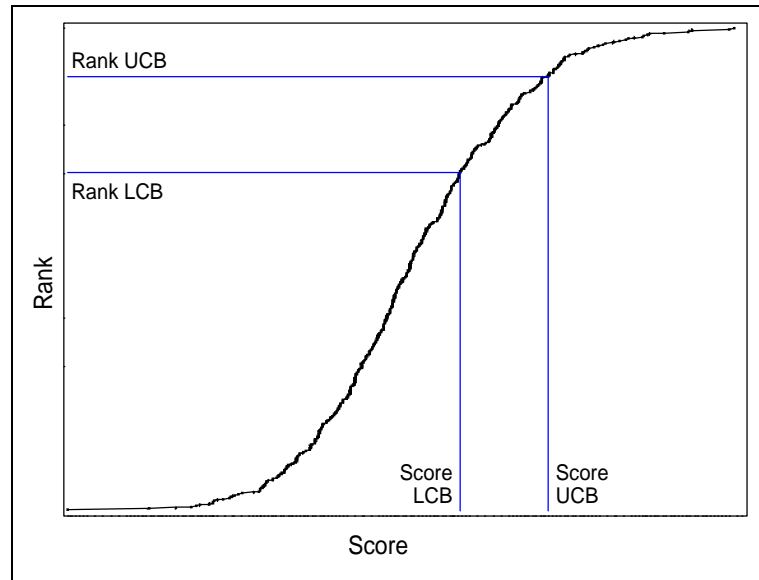
corresponding to this estimate is 0, a not unlikely crash outcome, even for a carrier whose crash rate estimate is this high.

**Table 15**  
**Crash Rate Estimates with Upper and Lower 95 percent Prediction Bounds**  
**(From Rate-Based Poisson Regression)**

DOT#	Number of PUs	Poisson LCB-95	Poisson Predicted	Poisson UCB-95
610052	1	0.00	0.17	1.00
261450	1	0.00	0.43	2.00
621939	1	0.00	0.18	1.00
582129	1	0.00	0.17	1.00
647010	1	0.00	0.15	1.00
739048	1	0.00	0.15	1.00
489456	1	0.00	0.17	1.00
279566	1	0.00	0.16	1.00
462126	1	0.00	0.16	1.00
659259	1	0.00	2.82	6.00
535838	1	0.00	2.78	6.00
...				

**Confidence Intervals for Ranks.** In addition to carrier risk or scores, percentile ranks can be computed from them, and it is actually the percentile ranks of carrier risk scores (among the scores for all carriers) that are used to classify them as “at risk,” “other identified,” and so on. Therefore, confidence ranges for ranks would also be useful to qualify carrier risk percentile ranks.

Figure 5 illustrates how to get a CI for a percentile rank from the CI for the corresponding carrier risk score. Associated with every carrier risk score is its rank among the multiple carrier risk scores computed in the regression. The function that returns the rank for a given score is plotted in Figure 5. Carrier risk score lower and upper confidence bounds (“Score LCB” and “Score UCB”) are shown on the figure’s horizontal axis. The rank function relates any score to a corresponding rank. Thus, lower and upper confidence bounds for the percentile ranks (“Rank LCB” and “Rank UCB”) are shown on the figure’s vertical axis.



**Figure 5: Confidence Bounds for a Hypothetical Score and its Corresponding Rank.**

As an example, consider again Table 15. Carrier 261450 (second row) has a predicted value of 0.43 crashes per PU and a confidence range of 0 to 2. The scores 0 and 2 correspond to ranks 1 and 117,443 (of 118,757 ranked carriers). Thus the confidence range 0 to 2 corresponds to percentile ranks of .001 percent (1/118,757) and 98.9 percent, which is nearly the whole range of percentiles. In view of these confidence limits, the point estimate of .43 crashes per PU is clearly no more than a rough approximation.

### 3.3 AN EMPIRICAL BAYES APPROACH

Empirical Bayes (EB) approaches to risk estimation combine risk estimates for individuals with risk estimates for a group at large. Because estimates for the group are statistically more stable than individual estimates, combined estimates can be more stable and better than estimates based on data for individuals alone. The EB estimates can also be adjusted for regression variables such as SEAs. The use of EB methods has become more and more standard in traffic and highway risk estimation (see for example 9, 10, 11). It therefore seems reasonable that EB estimators should be considered for determining baseline effectiveness in the SafeStat classification problem—among various reasonably standard alternative classification procedures.

**Empirical Bayes Estimators Defined.** Hauer (9) discusses EB estimates of crash rates for highway segments. The problem Hauer considers is directly analogous to the problem of estimating carrier risks. The number of crashes during a specified period is given by  $\mu \times L \times Y$ , where  $\mu$  is the number of crashes/km-year for a road segment,  $L$  is the road segment length, and  $Y$  is the length of the period of observation. Time and segment length are measures of exposure. In the SafeStat carrier risk estimation problem, time and the number

of power units are measures of exposure. The number of power units corresponds to segment length  $L$ .

The EB estimate Hauer considers is based on the following statistical model: Let  $\Gamma$  be a random variable having a gamma distribution, and suppose that given  $\Gamma=\gamma$ ,  $Y$  is a Poisson random variable with mean  $\gamma$ . Then it can be shown that the unconditional (i.e., over all possible values of  $\Gamma$ ) distribution of  $Y$  is negative binomial. It can also be shown that, for a given  $Y=y$ , the conditional distribution of  $\Gamma$  has mean  $w \times \mu + (1-w) \times y$ , where  $\mu$  is the mean of  $\Gamma$ ,  $\kappa$  is  $\mu^{-1}$  times the variance of  $\Gamma$ , and  $w = 1/(1+\mu/\kappa)$ . The parameter  $\kappa$  is called the dispersion parameter of the negative binomial distribution. The mean of the gamma distribution  $\mu$  is also (it can be shown) the mean of the negative binomial distribution.

The EB estimate is computed by estimating  $\mu$  and  $\kappa$  and hence  $w$  from group count data—multiple observed counts  $Y$  (for different road segments or carriers). Plugging estimates of  $\mu$  and  $w$  into the expression  $w \times \mu + (1-w) \times y$  for the conditional (given  $y$ ) mean crash count gives an estimate for the mean crash count for the individual with count  $y$ . This estimate is a weighted average of the estimate  $y$  for the individual and the estimate of  $\mu$  for the group.

Regression can be used to adjust the group mean estimates for the SEAs (or other predictor variables) for each individual. The adjusted group mean estimates and negative binomial dispersion parameter ( $\kappa$ ) can all be estimated with the SAS Genmod procedure. Because the  $\mu$  and  $\kappa$  are estimated, the EB approach circumvents assumptions about particular values of  $\mu$  and  $\kappa$ —assumptions that a classical Bayes approach would entail and that people sometimes find unsubstantiated (a problem with classical Bayes procedures).

**Application to SafeStat Data.** The negative-binomial EB estimator was also considered as a possible baseline alternative for the SafeStat risk classification problem. Using the SAS Genmod procedure, a negative binomial model analogous to the Poisson regression model considered earlier was fit to the SafeStat data. The regression model was the same; only the distribution (negative binomial instead of Poisson) was different. The SEA parameter estimates in the two cases turn out to be very similar. The estimate of the negative binomial model dispersion parameter  $\kappa$  is 1.8597.

The negative binomial regression approach, shown in Tables 16, 18, and 19, did better than the SafeStat approach (Tables 9 and 13) in all cases. Also, the negative binomial EB method (Tables 17, 20, 21) did better than the negative binomial regression approach in all cases. The negative binomial regression approach (Tables 16, 18, 19) was also better than the Poisson regression approach (Tables 10 and 13) at identifying carriers in the “All” and A&B groups, but the Poisson regression approach was better at identifying the “Other Identified (C)” carriers. The Poisson regression approach (Table 10) also did better than the EB approach (Tables 17) for the “Other Identified (C)” carriers. The EB approach (Table 17) did better than the LCB approach (Table 11). These results show that negative binomial EB estimation provides yet another reasonably standard approach to classification that outperforms the SafeStat method, generally by a substantial margin.

**Table 16**  
**Post-Selection Crash Rates**  
**for Negative-Binomial Model Carrier Groups**

<b>Carrier Groups Based on Negative Binomial Model</b>	<b>Carriers</b>	<b>Weighted Crashes</b>	<b>Power Units</b>	<b>Weighted Crashes per 1000 PUs</b>	<b>Percent Higher Than Not Identified</b>
All	118,757	58,607.8	2,293,655	25.55	5.9
All Identified	5,952	5,258.8	81,776	64.31	166.6
At-Risk (A&B)	3,595	2,132.3	28,655	74.41	208.5
Other Identified (C)	2,357	3,126.5	53,121	58.86	144.0
Not Identified	112,805	53,349.0	2,211,879	24.12	0.0

**Table 17**  
**Post-Selection Crash Rates**  
**for Empirical Bayes Carrier Groups**

<b>Carrier Groups Based on Empirical Bayes</b>	<b>Carriers</b>	<b>Weighted Crashes</b>	<b>Power Units</b>	<b>Weighted Crashes per 1000 PUs</b>	<b>Percent Higher Than Not Identified</b>
All	118,757	58,607.8	2,293,655	25.55	5.0
All Identified	5,952	4,383.0	65,226	67.20	176.2
At-Risk (A&B)	3,595	2,059.0	26,592	77.43	218.2
Other Identified (C)	2,357	2,324.0	38,634	60.15	147.2
Not Identified	112,805	54,224.8	2,228,429	24.33	0.0

**Table 18**  
**Negative Binomial Classifications for At-Risk Carriers**  
**Volpe Outliers Excluded, Sample Size = 3,595, Lift = 2.52**

	Post-Selection Crash Rate is High?		
Carrier Flagged by Method?	No	Yes	Percent Flagged
No	111,841	3,321	2.88
Yes	3,321	274	7.62

**Table 19**  
**Negative Binomial Classification for All Identified Carriers**  
**Volpe Outliers Excluded, Sample Size = 5,952, Lift = 2.22**

	Post-Selection Crash Rate is High?		
Carrier Flagged by Method?	No	Yes	Percent Flagged
No	107,514	5,291	4.69
Yes	5,291	661	11.11

**Table 20**  
**Empirical Bayes Classification for At-Risk Carriers**  
**Volpe Outliers Excluded, Sample Size = 3,595, Lift = 2.56**

	Post-Selection Crash Rate is High?		
Carrier Flagged by Method?	No	Yes	Percent Flagged
No	111,846	3,316	2.88
Yes	3,316	279	7.76

**Table 21**  
**Empirical Bayes Classification for All Identified Carriers**  
**Volpe Outliers Excluded, Sample Size = 5,952, Lift = 2.34**

	Post-Selection Crash Rate is High?		
Carrier Flagged by Method?	No	Yes	Percent Flagged
No	107,551	5,254	4.66
Yes	5,254	698	11.73

For purposes of comparison, the statistical methods have been constrained to select the same number of carriers as the SafeStat algorithm. In actual application, a control limit approach would be used to select only those carriers that exceed the expected value at some pre-determined probability level.

**Estimation vs Classification.** No optimality in the classification problem is assumed here for the EB negative binomial risk estimates. In the carrier risk classification problem, the express objective is to identify high risk carriers, not to estimate crash risks. Even if one estimator is better than another according to an estimation criterion such as mean square error, the second estimator might still correctly identify more high-risk carriers than the first. A best estimator is not necessarily a best classifier.

Of course a Bayesian perspective can also be taken in the classification problem, though such approaches are perhaps not as standard as the negative binomial EB approach considered above. Consider the two-by-two classification table.

Test Result	Actual Risk	
	Low (L)	High (H)
Not Identified (-)	A	B
Identified (+)	C	D
Total	A+C	B+D

A very general “prior” distribution is given by the probabilities of low and high-risk subjects:  $P(L)=A+C/(A+B+C+D)$  and  $P(H)=(B+D)/(A+B+C+D)$ . It is very general because it applies to all individuals in the low (L) or high-risk (H) groups, regardless of SEAs or other variables.

The conditional probabilities of the test results (+ or -) for each prior possibility (H or L) are  $P(+|H) = D/(B+D)$ ,  $P(+|L) = C/(A+C)$ ,  $P(-|H) = B/(B+D)$  and  $P(-|L) = A/(A+C)$ . By Bayes Theorem, the posterior probability (i.e., after assigning a + or -) of H given + is

$$P(H|+) = P(+|H)*P(H) / (P(+|H)*P(H) + P(+|L)*P(L)).$$

Similar expressions can also be written for the other posterior probabilities  $P(L|+)$ ,  $P(H|-)$ , and  $P(L|-)$ . When the prior is known—as in an effectiveness study—the (true) posterior probabilities can also be used to evaluate classification methods. For example  $P(H|+)/P(H)$  is the lift, defined previously.

This approach is sometimes called Bayes updating. Unless the assumed prior is specific to individuals, however, Bayes updating is not likely to be as powerful as the above negative binomial EB method with its carrier-specific estimates. No carrier-specific Bayes or EB approach expressly for the classification problem seems to be as standard as the negative binomial method. Therefore the negative binomial EB method is the only baseline EB approach that was implemented for this review.

## 4 DISCUSSION

The SafeStat system has been used by FMCSA to address motor carrier safety for several years now. The focus of this section is to address the specific issues listed below that were not addressed by the Volpe evaluation. The objective is to identify potential improvements to SafeStat.

### Data Issues

#### 1. How does the ongoing enforcement activity affect the updated evaluation?

The OIG report (1) supported the methodology of the original Volpe evaluation (2) noting that it was feasible because there was historical safety data available for analysis that was not influenced by decisions on the compliance review assignments. Now the SafeStat rankings play an integral role in motor carrier enforcement. Carriers ranked in March 2001 were subsequently targeted to improve their safety performance. The issue is the impact of the ongoing SafeStat system on carrier crash risk in the subsequent 18 months and, hence, the updated evaluation.

Volpe has responded in the updated evaluation (3):

“The effectiveness of these SafeStat-influenced safety programs in reducing crashes does to some extent mute the effectiveness of SafeStat to identify high crash risk carriers as measured in this study. The effectiveness of the SafeStat-influenced programs is significant. According to the a longitudinal study (12) conducted over a similar timeframe as the new effectiveness study, Category A and B carriers reduced their crash rate by 45 percent over a period of 18 months after they were identified in SafeStat. Repercussions of these programs on SafeStat-targeted carriers can lead to companies leaving the industry. The results of the new study showed that 21 percent of at-risk carriers (and 14 percent of the other identified carriers) were no longer actively operating interstate commercial motor vehicles after the 18 months from the March 2001 SafeStat identification. These numbers are significantly higher than the attrition rate of 5 percent of the carriers not identified by SafeStat.

Despite these complications, the new effectiveness study is still showing that SafeStat does work. The individual parts of SafeStat and SafeStat as a whole do identify carriers that are likely to have significantly higher crash rates than carriers not identified.”

About 10 percent of the carriers identified as at-risk in the March 2001 simulated SafeStat ranking are also high-risk based on their post-selection crash rate. There are two

interpretations of this result. One is that the SafeStat system compelled 90 percent of the at-risk carriers to improve their safety in the post-selection period. The other interpretation is that 90 percent were selected based on random variation in the data and regression to the mean is responsible for the reduced rates in the post-selection period. Since the original evaluation conducted before SafeStat was implemented produced similar results, the later explanation is more plausible.

## **2. What is the magnitude and impact of missing data?**

Based on available national data, underreporting by the states to the Crash file can be estimated at approximately one-third of the expected national total, or 50 percent of the known cases. While state reporting of fatal crashes is nearly complete, only about 80 percent of injury and 50 percent of tow-away crashes are reported. Geographic variations in missing data are likely to bias the SafeStat rankings.

## **3. What is the magnitude and impact of late data?**

For the 30-month SafeStat window, late data are estimated to reduce the number of crashes by about 25 percent. As with missing data, there are geographic variations in late data that have the potential to bias the SafeStat rankings. The original March 2001 ranking was compared with the simulated ranking that included all late data received as of March 2003. Overall, the SafeStat category was unchanged for 81 percent of the carriers. The number of at-risk carriers (A-B) was increased by 33 percent by the movement of carriers previously ranked lower. However, this increase was offset by the movement of 15 percent of the at-risk carriers to lower rankings, for a net increase in at-risk carriers of 18 percent. Incomplete data resulted in some carriers being ranked at-risk when they would not have been with complete data.

## **4. Can the algorithm be modified to adjust for missing data?**

State-to state patterns of underreporting were examined from year to year. While some state problems have persisted, the more typical situation is for the underreporting to shift from state-to-state and year-to-year as known problems are addressed and new ones arise. This situation prevents model adjustments that would compensate for missing data.

## **Formula Issues**

## **5. What has been the impact of formula changes?**

ORNL did not have the necessary source data to evaluate formula changes.

## **6. Do the current formula weights provide the best indication of carrier safety?**

The formula weights have been established on the basis of expert opinion. First, the indicators are combined to produce the SEAs and then the overall score is a weighted combination of the four SEAs. For example, the accident, driver, vehicle, and safety

management SEAs are weighted 2.0, 1.5, 1.0,1 respectively. Statistical models confirm the association of the SEAs to crash risk, although the relative weights put more emphasis on crashes.

Here are the time and severity weighting of crashes, taken from the Volpe documentation in Appendix A:

**Time Weights ( $W_t$ )**

For post crashes within:	Weight:
0 to 6 months after 24-Mar-01	1.5
7 to 12 months after 24-Mar-01	1.0
13 to 18 months after 24-Mar-01	0.5

**Severity Weights ( $W_s$ )**

For post crashes with:	Weight:
Only a tow-away	0.5
(Injury or Fatality) and (No HM Release)	1.0
(No Injury nor Fatality) and (HM Release)	1.0
(Injury or Fatality) and (HM Release)	1.5

The overall weight for each crash is computed by multiplying the time weight with the severity weight:

$$W = W_t * W_s$$

While the rationale for this weighting is apparent, the net effect is to reduce the crash count. This probably occurs because about half of the crashes are tow-away with a weight of 0.5. Further, rare events have greater variance than more frequent events. Giving rare events, such as injury/fatality with HM release, larger weights increases the overall variability of the resulting measure. If rare events are important to the ranking, they should be ranked separately so that the variance of these events is given appropriate consideration. The information in the data is misrepresented when observations are weighted. The current weighting in the SafeStat algorithm is probably counter productive.

**Data Sufficiency**

A carrier is not ranked unless specified data sufficiency requirements are met. These requirements were taken from the documentation provided by Volpe in Appendix A.

- (1) A carrier had a CR within 18 months of the simulation date of 23-Mar-01 or
- (2) A carrier met 2 or more of the following conditions:
  - a. Had 3 or more driver inspections in the last 30 months
  - b. Had 3 or more vehicle inspections in the last 30 months
  - c. Had 3 or more HM OOS inspections in the last 30 months

- d. Had an enforcement in the last 6 years
- e. Had 2 or more state-reported crashes in the last 30 months.

In addition, each of the carriers must have had an “active” status as of end of the post-simulation crash period and had non-zero power unit values at that time in order to be included in the effectiveness study.

These sufficiency requirements are also counter productive for a statistical model. Of the 118,757 carriers used for the Volpe evaluation, only 17,157 have 2 or more crashes and of the 555,259 active carriers, only 19,430 have more than 2 crashes. In the post-selection period, one crash was recorded for about 60 percent of all carriers with an crash count greater than zero. Omitting this much data distorts the distribution and increases the unexplained variation. The same is true for all the other data counts.

A statistical model will be improved by including all the data. Coefficients (weights) for the various indicators are objectively determined by the relationship of each factor to crash risk in the historical data rather than by expert judgment. The coefficients in the models developed for this study are different than the corresponding SafeStat weights, although each of the SEAs was significantly associated with crash risk. Since the statistical model provides estimates of the variance of the modeled result, the affect of small samples is taken into account. The SafeStat algorithm omits more data than it uses but this is not sufficient to prevent the selection of many carriers that are not high-risk in the post-selection period.

**7. Would vehicle miles traveled (vmt) or an adjustment to fleet size for team drivers provide an improved basis for normalizing (exposure) than the current average number of power units?**

The choice of an exposure measure to normalize the counts of violations and crashes is a difficult one. Truck crash rates are known to vary substantially from one road type to another. Except for the peer groups, all carriers are held to the same standard in the SafeStat algorithm. Dividing by either number of power units or vmt limits the method to a linear relationship. Statistical models provide more alternatives by moving the exposure variable to the other side of the equation with the other independent (predictor) variables. A non-linear function is likely to yield the best fit and such a function would compensate for higher, or safer, travel by larger fleets. This would accomplish something similar to the peer groups in that carriers would essentially be compared with other carriers of comparable size, except that the coefficients would be determined by the modeling process.

**8. Are the rankings consistent?**

The Volpe evaluation looks at the aggregate experience of at-risk carriers versus the remaining carriers. But the rankings impact individual carriers, so it is also important to evaluate the consistency of the ranking for individual carriers. The OIG report points out that the at-risk threshold varies among the peer groups. Frequently, the number of events (crashes or violations) that determine a carrier’s rank are quite small. Any statistic (calculated quantity) is subject to the variability, or error, in the source data. This variability

is a combination of random error and bias error. Can a carrier be classified as at-risk due to random variability or bias?

Another way to evaluate the SafeStat rankings is to use the post-selection crash rates to rank the carriers and compare this ranking to the original SafeStat ranks. The advantage of this approach is that it shows the post-selection outcome for each carrier, rather than the aggregate experience. This method of evaluation shows that 90 percent of the carriers identified by the Volpe SafeStat algorithm did not have a high crash risk in the post-selection period.

This result is a reflection of the large variability inherent in crashes and crash rates. Apparently, 90 percent of the carriers identified by the SafeStat algorithm as at-risk are classified due to random variations and are not consistently high risk carriers. This situation is referred to as “regression to the mean.” Regression to the mean occurs anytime one selects the extreme values on a measure that includes random variations. The SafeStat algorithm addresses variability in a qualitative way. A carrier is not ranked unless specified data sufficiency requirements are met. This review shows that these procedures are not sufficient. The inherent variability in the data limit the inferences that can be made. The SafeStat algorithm produces a ranking, but it does not acknowledge the limitations in this result. This is the most serious shortcoming. Statistical methods have the advantage that the variability is quantified. A ranking, or measure, based on a statistical model will also estimate the variability in the measure that arises from the variability in the source data.

## 5 CONCLUSIONS

This review addressed the magnitude and impact of missing and late crash data and issues related to the SafeStat algorithm. Conclusions based on the assessment of missing and late crash data are summarized below.

1. Underreporting to the MCMIS Crash file is estimated to be about one-third.
  - Reporting of fatal crashes is nearly complete
  - About 80 percent of injury crashes are reported
  - About 50 percent of non-injury crashes are reported
2. It takes about 16 months for the MCMIS Crash file to receive 90 percent of the crash reports from the states. This lag removes an additional 25 percent of the crash reports from the SafeStat window of 30 months.
3. Late data has an impact on the SafeStat rankings. Overall, the rankings of 81 percent of the carriers were unchanged with the addition of late data. The number of at-risk carriers is increased by 33 percent by the movement of carriers previously ranked lower. However, this increase is offset by the movement of 15 percent of the at-risk carriers to lower rankings, for a net increase in at-risk carriers of 18 percent. Incomplete data resulted in some carriers being ranked at-risk when they would not have been with complete data. Since the Volpe evaluation used historical data, this impact of late data is not reflected in the Volpe evaluation. Missing data may have a similar impact.
4. Geographic and year-to-year variations in missing and late data are likely to bias the SafeStat rankings.

Statistical methods applied in this review demonstrate some shortcomings in the Safestat algorithm that were not identified in the Volpe evaluation. These conclusions are summarized here.

5. The ORNL review finds that the SafeStat algorithm is about twice as effective as random selection in identifying high-risk carriers. This is consistent with the Volpe finding that the aggregate crash rate for at-risk carriers is about double the rate for carriers not identified as high-risk.
6. Statistical models demonstrate the association of the four SEAs to crash risk, although the relative weighting differs from SafeStat.
7. All of the statistical models were more effective than the Safestat algorithm, although the improvement was modest (30 percent). The use of unweighted and uncensored

data may improve the effectiveness of the statistical methods.

8. The Volpe evaluation is based on the aggregate crash rates and does not address the consistency of the rankings for individual carriers. The ORNL review finds that about 90 percent of carriers identified as at-risk did not have a high crash risk in the post-selection period.
9. The ORNL review concludes that most carriers are identified as at-risk by SafeStat due to random variations in the source data rather than a significant change in carrier risk. The Volpe SafeStat algorithm does not adequately take into account the inherent variability in the scores when identifying high risk carriers. This leads to the selection of some carriers as high risk when their score does not exceed the inherent variability of the SafeStat scores. Selecting carriers with the highest scores, without considering the accuracy of the scores, results in the selection of many carriers due to random variations and not any significant change in carrier risk. In this situation, random variations would also be expected to return these carriers to expected risk levels in a subsequent observation period. This phenomenon is referred to as “regression to the mean.” Provisions in the Safestat algorithm to assure sufficient information are not adequate.

Statistical methods are available to quantify the variability inherent to the data and correct for regression to the mean. The application of statistical methods can distinguish carriers with significantly elevated safety risk from those with risk levels that do not exceed the variability in the source data. For example, empirical Bayes approaches are now widely accepted for the very similar problem of selecting highway sites for treatment (9). Statistical models can be used to select coefficients (weights) for the various measures based on the relationship to collision risk in the historical data. This approach would replace expert judgment with objective statistical methods. While statistical methods can quantify random variations, they cannot correct for bias error. Improving the timeliness and completeness of the source data are still essential.

## REFERENCES

- 1 U.S. Department of Transportation Office of the Inspector General. *Improvements Needed in the Motor Carrier Safety Status Measurement System*. February 2004.
- 2 Madsen, D.G and Wright, D.G. *An Effectiveness Analysis of SafeStat*. TRB paper No. 990448. November 1998.
- 3 Volpe National Transportation Systems Center. *SafeStat Effectiveness Study Update*. March 2004.
- 4 Volpe National Transportation Systems Center. *Overview Report of State Safety Data Quality*. June 2004. <http://ai.volpe.dot.gov/SafeStat/document/Overview.htm>
- 5 Blower, D. and Matteson, A. *Evaluation of the Motor Carrier management Information System Crash File, Phase I*. University of Michigan Transportation Research Institute. UMTRI-2003-6. March 2003.
- 6 Blower, D. and Matteson, A. *Evaluation of Missouri Crash Data Reported to MCMIS Crash File*. University of Michigan Transportation Research Institute. UMTRI-2004-5. January 2004.
- 7 Blower, D. and Matteson, A. *Patterns of MCMIS Crash File Underreporting in Ohio*. University of Michigan Transportation Research Institute. August 2003.
- 8 Federal Motor Carrier Safety Administration. *Motor Carrier Safety Progress Report*. March 2004.
- 9 Hauer, E., D.W. Harwood, F.M. Council, M.S. Griffith. *The Empirical Bayes Method For Estimating Safety: A Tutorial*. Transportation Research Record 1784, pp. 126-131. National Academies Press, Washington, D.C. (2002).
- 10 Hauer, E. "Empirical Bayes approach to the estimation of "unsafety": The multivariate regression method." *Accident Analysis and Prevention*, Vol. 24, No. 5, 457-477, 1992.
- 11 Persaud, B.N., R.A. Retting, P.E. Garder, and D. Lord. *"Observational Before-After Study of the Safety Effect of U.S. Roundabout Conversions Using the Empirical Bayes Method*. presented at the 80th Annual Meeting of the Transportation Research Board, January 2001, Preprint CD-ROM, ID# 01-0562.
- 12 Volpe National Transportation Systems Center. *Measuring the FMCSA's Safety Objectives from Year 2000 to 2002*, July 2003.

## APPENDIX A:

### Additional Notes Provided by Volpe on the Data Files

#### SafeStat Effectiveness Study Update

(3/15/04)

#### Introduction

The intention of the SafeStat Algorithm is to identify high safety risk carriers. The design and development of SafeStat began in 1994 and the first nationwide results were produced in 1997. The purpose of the updated SafeStat Effectiveness Study is to confirm SafeStat's safety risk identification capabilities with new data (MCMIS data available as of March 2003 and new version of SafeStat methodology, version 8.5).

The effectiveness study first produces a simulated SafeStat run with a start date of 24-Mar-01 under a set of specific constraints. Then crashes that occurred up to 18 months after the simulation for those same carriers in the simulation are collected; these are referred to as 'post-simulation crashes'. Crash rates among subsets of the carrier population are used to assess the SafeStat algorithm.

#### Simulation Algorithm and Data

The results of the SafeStat Simulation are in a SPSS format and named **Sea01\_Simulation2.sav**. An associated data dictionary is placed in the appendix of this document. The simulation employed SafeStat algorithm version 8.5. It was determined that the 28-Sep-02 power unit data were more accurate than those of 24-Mar-01, as a consequence of FMCSA's industry requirement imperative on MCS-150 updates. Therefore, the Sep-02 power units were used in the simulation for both the simulated SafeStat run and the post simulation crash period. The average power unit calculation used in the Accident Involvement Indicator (AII) in version 8.5 was not employed. Also, only carriers that were resident in the 50 United States and Washington D.C. were used in the simulation.

#### Post-Simulation Crash Data

The MCMIS data load of 24-Mar-2003 was used for obtaining the post-simulation crash data. The crash data spanned a period from 25-Mar-2001 to 24-Sep-2002. Only crashes for carriers that were included in the simulation were used. Power unit values computed as of Sep-2002 were employed to compute the crash rates.

#### Time and Severity Weighted Post-Simulation Crashes

In this study, each subpopulation in the simulation is evaluated based on the crash involvement over the 18-month post simulation period. The evaluation procedure applies weights according to the severity of crash as well as the time delay between the crash and the time of the simulation. The weight schedule is shown below:

Time Weights ( $W_t$ )

For post crashes within:	Weight:
0 to 6 months after 24-Mar-01	1.5
7 to 12 months after 24-Mar-01	1.0
13 to 18 months after 24-Mar-01	0.5

#### Severity Weights ( $W_s$ )

For post crashes with:	Weight:
Only a tow-away	0.5
(Injury or Fatality) and (No HM Release)	1.0
(No Injury nor Fatality) and (HM Release)	1.0
(Injury or Fatality) and (HM Release)	1.5

The overall weight for each crash is computed by multiplying the time weight with the severity weight:

$$W = W_t * W_s$$

#### The Effectiveness Process

The SafeStat Simulation results in conjunction with the post-simulation crash data are used as input to a program that produces a SafeStat effectiveness result file in SPSS data format named **ss\_eff02.sav**. The SafeStat effectiveness process selects the following carriers with sufficient data to receive and assessment by SafeStat. The criteria for data sufficiency are as follows:

- (3) A carrier had a CR within 18 months of the simulation date of 23-Mar-01 or
- (4) A carrier met 2 or more of the following conditions:
  - a. Had 3 or more driver inspections in the last 30 months
  - b. Had 3 or more vehicle inspections in the last 30 months
  - c. Had 3 or more HM OOS inspections in the last 30 months
  - d. Had an enforcement in the last 6 years
  - e. Had 2 or more state-reported crashes in the last 30 months.

In addition, each of the carriers must have had an “active” status as of end of the post-simulation crash period and had non-zero power unit values at that time in order to be included in the effectiveness study.

Finally, the SafeStat Effectiveness Study included an effort to identify outliers. Carriers with very high crash rates as well as very low crash rates were found. The outlier identification process is described below.

#### Filtering Outlier Carriers with Suspect Post Simulation Crash Rates

The November 1998 Effectiveness Study employed a Poisson-distributed based filtering procedure to identify outliers with suspiciously high post simulation crash rates and the same parameters were also applied to this study. This procedure identified high crash rate carriers that were excluded from the updated effectiveness study results. However, two carriers,

Prime Inc (DOT# = 3706) and Dick Simon (DOT# = 21331), identified as having suspiciously high crash rates by the filter, were included in the study based on knowing that there their high crash rates are function of their safety and not data accuracies.

Criteria were also applied to identify carriers with suspiciously low crash rates during the post simulation crash period. Carriers with one or more crash and a crash rate of 1 per 1000 power units (PUs) or lower during the post simulation crash period were excluded from the study results. Carriers with a crash rates greater than 1 crash per 1000 PUs and less than 4.5 crashes per 1000 PUs were removed if there were other collaborating evidence that the number of PUs was not indicative of the carrier's exposure over the study time frame. These carrier were removed from the study results if:

- (1) the power units decreased by more than 2/3 during the period from Sept-02 to Oct-03 had very high power units, or
- (2) the vehicle or driver inspection rate (insp. / power units) was less than .02 during the Sept-98 to Mar-01 period

Carriers with no crashes during the post-simulation crash period were removed from the study results if they had 800 or more PUs. Carriers with no crashes and 233 to 799 PUs were removed if there were other collaborating evidence that the number of PUs was not indicative of the carrier's exposure over the study time frame. These carriers were removed from the study results if:

- (1) the power units decreased by more than 2/3 during the period from Sept-02 to Sept-03, or
- (2) there were no inspections or a noticeable drop in inspections during the post-simulation period.

Three other carriers that primarily lease vehicles to other carriers were identified as outliers. These three carriers had large power unit fleets, but had small numbers of drivers. The crashes and violations should be assigned to the carriers that leased the vehicles, but often the leasing company incorrectly absorbs the statistics. This leads to extreme SafeStat results and post-simulation crash rates that are most likely not indicative of the actual carrier performance.

## Files

There are three files in SPSS format:

- (1) **SS\_rslt12.sav**: Contains the actual March 24, 2001 SafeStat (version 6.1) results produced by the FMCSA on that date. These results included 583,438 U.S., Canadian and Mexican carriers. See listing that follows:
- (2) **Sea01\_Simulation2.sav**: Contains the results of the simulated SafeStat run for the 555,259 carriers. This file includes the all of the carriers domiciled in the U.S. and their SafeStat results from the simulated SafeStat run. See listing that follows:
- (3) **ss\_eff02.sav**: Contains the results the 118,907 carriers that had sufficient data to be used in the study. This includes the carriers used in the final results and the outliers. Both Simulation SafeStat and post-simulation crash rate information are contained in this file. See listing that follows:

Data Dictionary for fields in **SS\_rslt12.sav**

- Original FMCSA SafeStat Results for March 2001 (583,438 Carriers)

DOT#	-U.S. DOT Number of the carrier
STATE_CO	-State of Residence
REGION	Region
OIC_CARR	-OIC_CARRIER
CARRIER_	-Carrier Name
SCORED_L	-Carrier received a SS Score in the last run
V7	- Carrier City
V8	- Carrier County
HM_PASS	-Hazmat / Passenger Flag
POWER_UN	-Number of Power Units based on the March 2001 MCMIS snapshot
CENSUS_R	-CENSUS_REVIEW_DATE
V12	-CENSUS_RATING
NBR_DRIV	-Number of Drivers
NBR_INTE	-Number of Interstate Drivers
NBR_INTR	-Number of Intrastate Drivers
WARNING_	-Warning Letter Count
V17	-Warning Letter Date
RUN_DATE	-Run Date Century
V19	-Run Date Year
V20	-Run Date Month
V21	-Run Date Day
CURR_MST	-Current MCMIS Step Date Century
V27	-Current MCMIS Step Date Year
V28	-Current MCMIS Step Date Month
V29	-Current MCMIS Step Date Day
VIM	-Vehicle Inspection Measure
TMPVII	-Temporary Vehicle Inspection Indicator
VII	-Vehicle Inspection Indicator
VHGRP	-carriers peer grouping by number of vehicle inspections
VHINCF	-vehicle inspection data confidence level
DIM	-Driver Inspection Measure
TMPDII	-Temporary Driver Inspection Indicator
DII	-Driver Inspection Indicator
DRGRP	- carriers peer grouping by number of driver inspections
HMIM	-Hazmat Inspection Measure (not currently used)
TMPHMII	-temporary hazmat inspection indicator (not currently used)
HMII	-Hazmat Inspection Indicator (not currently used)
HMGRP	- carriers grouping by number of hazmat inspections (not currently used)
TT_DO	-total driver OOS inspections (last 30 months)
TT_DV	-total driver OOS violations (last 30 months)
TT_VO	-total vehicle OOS inspections (last 30 months)
TT_VV	-total vehicle OOS violations (last 30 months)
TT_HMO	-total hazmat OOS inspections (last 30 months) (not currently used)
TT_HMV	-total hazmat OOS violations (last 30 months) (not currently used)
DRINCF	-driver inspection data confidence level
TW_DO	-total weighted driver OOS inspections (last 30 months)
TW_DV	-total weighted driver OOS violations (last 30 months)
TW_DRIN	-total weighted driver inspections (last 30 months)
TW_VO	-total weighted vehicle OOS inspections (last 30 months)
TW_VV	-total weighted vehicle OOS violations (last 30 months)
TW_VHIN	-total weighted vehicle inspections (last 30 months)

TW_HMO	-total weighted hazmat OOS inspections (last 30 months)
TW_HMV	-total weighted hazmat OOS violations (last 30 months)
TW_INSP	-total weighted inspections (last 30 months)
ADR	-actual driver OOS rate
ADVR	-actual driver OOS violation rate
AVR	-actual vehicle OOS rate
AVVR	-actual vehicle OOS violation rate
NO_OOSV	- number of inspections in which there violations to an OOS order
NBR_VHIN	-total number of vehicle inspections (last 30 months)
NBR_DRIN	-total number of driver inspections (last 30 months)
MVGRP	-moving violation peer group
TW_MV	-number of weighted moving violations in the last 30 months
TT_MV	-number of moving violations in the last 30 months
TMPMVI	-temporary moving violation indicator
MVI	- moving violation indicator
MVM	- moving violation measure
REVIEW_T	-review type (C=Compliance, H=Hazmat, etc)
FACTOR_1	-general factor (S=satisfactory, C=conditional, U=unsatisfactory)
FACTOR_2	-driver factor (S=satisfactory, C=conditional, U=unsatisfactory)
FACTOR_3	-operational factor (S=satisfactory, C=conditional, U=unsatisfactory)
FACTOR_4	-vehicle factor (S=satisfactory, C=conditional, U=unsatisfactory)
FACTOR_5	-hazmat factor (S=satisfactory, C=conditional, U=unsatisfactory)
FACTOR_6	-accident factor (S=satisfactory, C=conditional, U=unsatisfactory)
OVERALL_	-Overall Review Rating (S=satisfactory, C=conditional, U=unsatisfactory)
LAST_REV	-Last Review Date
REV_FORM	-Review Form Number (a unique CR identifier)
RECORDAB	-Number of recordable crashes in the last year
REV_VIOL	-Review Violations Discovered
RAR	-Recordable Accident Rate
RAI	-Review accident indicator
TMPRAI	-Temporary Review accident indicator
RVCNFDRI	- driver review confidence level
RVCNFVRI	- vehicle review confidence level
REVIEW_A	-Review Age (months)
V93	-Review Age Weight
MILEAGE	-Vehicle Miles Traveled
RAGRP	-recordable accident peer group
FAC_S	-Number of factors rated as satisfactory
FAC_C	-Number of factors rated as conditional
FAC_U	-Number of factors rated as unsatisfactory
FAC_N	-Number of non-rated factors
DRM	-Driver review measure
VHRM	-Vehicle review measure
HMRM	-Hazmat review measure
SMRM	-Safety Management review measure
DR_RECS	-Driver review records
VH_RECS	-Vehicle review records
HM_RECS	-Hazmat review records
SM_RECS	-Safety Management review records
TMPDRI	-Temporary Driver Review indicator
DRI	- Driver Review indicator
TMPVRI	-Temporary Vehicle Review indicator
VRI	- Vehicle Review indicator
TMPSMRI	-Temporary Safety Management Review indicator
SMRI	- Safety Management Review indicator

TMPHMRI	-Temporary Hazmat Review indicator
HMRI	- Hazmat Review indicator
REVIEW	- X if there was a CR in the last 18 months
HM_REVIEW	- X if there was a Hazmat Review in the last 18 months
VIOL_HM_V120	-Number of Hazmat Acute Viol
VIOL_HM_V122	-Number of Hazmat Critical Viol
VIOL_DR_V122	-Number of Driver Acute Viol
VIOL_DR_V124	-Number of Driver Critical Viol
VIOL_VH_V124	-Number of Vehicle Acute Viol
VIOL_VH_V126	-Number of Vehicle Critical Viol
VIOL_SM_V126	-Number of Safety Management Acute Viol
VIOL_SM_V126	-Number of Safety Management Critical Viol
LAST_ENF	-Last Enforcement Date
ENF_CR_D	-Date of the last CR that followed the Enforcement
AC_VIOLS	-sum of the acute/critical violations in the most recent CR
ESM	- enforcement severity measure
PESM	-rank value of the esm employing percent subcommand and used to compute the ehi in group 2
EHI	-Enforcement History Indicator
NBR_OF_E	-Number of enforcements
ENFCNF	1 if ehi value is less that 75; 2 if ehi is 75 or greater
RECENT	-age of the most recent enforcement (months)
AIM	-accident involvement measure
TCTWA	-total NGA accident measure.
All	-Accident Involvement Indicator
NBR_ACCI	-number of state reported crashes
N_ACCINU	-number of state reported crashes since the last CR
TMPAll	-Temporary Accident Involvement Indicator
ACGRP	-state reported crash peer group
LAST_CRA	-age of the newest state reported crash (months)
ACSEA	-Accident SEA
DRSEA	-Driver SEA
SMSEA	-Safety Management SEA
VHSEA	-Vehicle SEA
ACSEA_MI	1 if the Accident SEA is missing
DRSEA_MI	1 if the Driver SEA is missing
SMSEA_MI	1 if the Safety Management SEA is missing
VHSEA_MI	1 if the Vehicle SEA is missing
SEA_COUN	-number of deficient SEAs
SAFESTAT	-SafeStat Score
TMPSCORE	-Temporary SafeStat Score
SEA_CATE	-SEA Category
SSRANK	-SafeStat RANK
SSRANK_S	-SafeStat State RANK
REC_MCSI	-REC_MCSIP_STEP
CURR_MCS	-Current MCSIP Step
STEP_DAT	-MCSIP Step Date
ISS_SUM	-ISS_SUM
ISS_GROU	-ISS_GROUP

Data Dictionary for fields in **Sea01\_Simulation2.sav**

– SafeStat Simulation used for the 2003 Effectiveness Study ( 555,259 Carriers)

dot# -U.S. DOT Number of the carrier

The next seven fields are referred to in the section above on: Filtering Outlier Carriers

lowbound 1 if Carrier has an extremely low crash rate; is an outlier  
 hi\_bound 1 if Carrier has an extremely high crash rate; is an outlier  
 lease 1 if Carrier leases trucks and does not generally operate; is an outlier  
 outlier 1 if Carrier is an outlier as a consequence of extreme crash rate or leases trucks  
 pu -Number of Power Units based on the September 2002 MCMIS snapshot  
 thrsh The threshold value. A crash count at or above this value indicates a very high crash rate  
 pow\_u The number of power units associated with the post-simulation period. If the carrier did not participate during this period, pow\_u will be missing.

The next eleven fields are referred to in the section above on: The Effectiveness Process

attrit 1 if the carrier did not participate in the post-simulation period; or else 0  
 no\_pu 1 if the power unit value is missing or zero; or else 0.  
 Outstudy A value of 1 indicates that the carrier will not be included in the study because there was nonzero power units or the carrier dropped out of the 18 month post-simulation period.  
 Chk\_a 1 indicates that the carrier had a CR before 18 months of 24-Mar-01. Carrier had sufficient data  
 Chk\_b 1 indicates that there were 3 or more driver inspections within 30 months of the simulation run.  
 Chk\_c 1 indicates that there were 3 or more vehicle inspections within 30 months of the simulation run.  
 Chk\_d 1 indicates that there were 3 or more HM OOS inspections within 30 months of the simulation run  
 Chk\_e 1 indicates there was at least one enforcement within 6 years of the simulation run  
 Chk\_f 1 indicates if there was 2 or more state-reported crashes  
 Types 1 indicates the carrier is in SEA Category A or B;  
 2 indicates the carrier is in SEA Category C  
 3 indicates the carrier has sufficient data as evidenced by:  
 either a CR before 18 months of 24-Mar-01 or  
 a value of 2 or more in Check field

carrier\_ -Carrier's Name  
 state\_co -State of Residence  
 omc\_regi -Resource Center  
 status -A if Company is active, is a carrier or carrier/shipper, is interstate or intrastate  
 hazmat

revcc revyr revmo revday  
 -Date of the most recent compliance review (century, year, month, day)

safety\_r -Safety Rating  
 mcsip\_st -MCSIP Step  
 mcsip\_da -MCSIP Date  
 entity\_t -Entity Type (B = both carrier/shipper; C = carrier only)  
 hazmat\_f -Hazmat Flag (Y=Carrier hauls hazmat; N=Carrier does not haul hazmat)  
 idcutdat -MCMIS snapshot date  
 tw\_hmo -total weighted hazmat OOS inspections (last 30 months)  
 tw\_hmv -total weighted hazmat OOS violations (last 30 months)  
 tw\_insp -total weighted inspections (last 30 months)

tt_insp	-total number of inspections (last 30 months)
tt_vhin	-total number of vehicle inspections (last 30 months)
tt_drin	-total number of driver inspections (last 30 months)
tt_do	-total driver OOS inspections (last 30 months)
tt_vo	-total vehicle OOS inspections (last 30 months)
tt_dv	-total driver OOS violations (last 30 months)
tt_vv	-total vehicle OOS violations (last 30 months)
tt_hmo	-total hazmat OOS inspections (last 30 months) (not currently used)
tt_hmv	-total hazmat OOS violations (last 30 months) (not currently used)
vim	-Vehicle Inspection Measure
avr	-actual vehicle OOS rate
avvr	-actual vehicle OOS violation rate
vhgrp	-carriers peer grouping by number of vehicle inspections
tmpvii	-temporary vehicle inspection indicator
vii	-vehicle inspection indicator
vhincnf	-vehicle inspection data confidence level
dim	-Driver Inspection Measure
adr	-actual driver OOS rate
advr	-actual driver OOS violation rate
n_jumps	-total number of violations to an OOS order
no_oosv	- number of inspections in which there violations to an OOS order
drgrp	- carriers peer grouping by number of driver inspections
tmpdii	-temporary driver inspection indicator
dii	-driver inspection indicator
drincnf	-driver inspection data confidence level
hmim	-Hazmat Inspection Measure (not currently used)
hmgrp	- carriers grouping by number of hazmat inspections (not currently used)
tmphmii	-temporary hazmat inspection indicator (not currently used)
hmii	-Hazmat Inspection Indicator (not currently used)
esm	- enforcement severity measure
all_cnt	-total number of enforcements in the last 6 years
esm_c	-enforcement severity measure – CR component
lstenfdt	-last enforcement date
lstenfag	-last enforcement age
lstr_yr, lstr_mo, lstr_day	-last enforcement date components (year, month, day)
cr_enf	-number of days between the last enforcement and the following CR
ac_viol	-number of acute/critical violations in the CR following the last enforcement
cr_cnt	-number of CR-related enforcements
esm_n	-enforcement severity measure – non-CR component
ncr_cnt	-number of non-CR-related enforcements
group	-if group=1 then ehi is set 75 or greater; if group=2 then ehi is set less than 75 (see SafeStat methodology document)
pesm	-rank value of the esm employing percent subcommand and used to compute the ehi in group 2
ehi	-Enforcement History Indicator
pesm2	-rank value of the esm employing percent subcommand and used to compute the ehi in group 1
revtype	-Review type (C = Compliance Review)
overall_	-Overall Review Rating (S=satisfactory, C=conditional, U=unsatisfactory)
factor_1	-general (S=satisfactory, C=conditional, U=unsatisfactory)
factor_2	-driver factor (S=satisfactory, C=conditional, U=unsatisfactory)
factor_3	-operational factor (S=satisfactory, C=conditional, U=unsatisfactory)

factor_4	-vehicle factor (S=satisfactory, C=conditional, U=unsatisfactory)
factor_5	-hazmat factor (S=satisfactory, C=conditional, U=unsatisfactory)
factor_6	-accident factor (S=satisfactory, C=conditional, U=unsatisfactory)
mcs15\$17	-vehicle miles traveled during the last year since the CR
pros_no	-prosecution number
revzz	-the year of the most recent CR
rev_age	-age of the most recent CR (in months)
tt_rate	-total number of ratable acute/critical violations in the CR
tt_nrate	-total number of non-ratable acute/critical violations in the CR
dr1_meas	-level 1 sum of the CR driver measures
dr2_meas	-level 2 sum of the CR driver measures
dr3_meas	-level 3 sum of the CR driver measures
hm1_meas	-level 1 sum of the CR hazmat measures
hm2_meas	-level 2 sum of the CR hazmat measures
hm3_meas	-level 3 sum of the CR hazmat measures
sm1_meas	-level 1 sum of the CR safety management measures
sm2_meas	-level 2 sum of the CR safety management measures
sm3_meas	-level 3 sum of the CR safety management measures
vh1_meas	-level 1 sum of the CR vehicle measures
vh2_meas	-level 2 sum of the CR vehicle measures
vh3_meas	-level 3 sum of the CR vehicle measures
dr1_recs	-level 1 sum of the CR driver records
dr2_recs	-level 2 sum of the CR driver records
dr3_recs	-level 3 sum of the CR driver records
hm1_recs	-level 1 sum of the CR hazmat records
hm2_recs	-level 2 sum of the CR hazmat records
hm3_recs	-level 3 sum of the CR hazmat records
sm1_recs	-level 1 sum of the CR safety management records
sm2_recs	-level 2 sum of the CR safety management records
sm3_recs	-level 3 sum of the CR safety management records
vh1_recs	-level 1 sum of the CR vehicle records
vh2_recs	-level 2 sum of the CR vehicle records
vh3_recs	-level 3 sum of the CR vehicle records
level_1	-level 1 sum of the CR records
level_2	-level 2 sum of the CR records
level_3	-level 3 sum of the CR records
ac_count	-sum of the acute/critical violations in the most recent CR
no_seas	-number of CR violations in the most recent CR
hmrn	-hazmat review measure
smrn	-safety management review measure
drn	-driver review measure
vrn	-vehicle review measure
ac_vbad	-earlier version of the acute/critical violation count (no longer used)
tmpmri	-temporary hazmat review indicator
tmpdri	-temporary driver review indicator
tmpvri	-temporary vehicle review indicator
tmpsmri	-temporary safety management review indicator
dr_recs	-sum of the CR driver records (all levels)
hm_recs	-sum of the CR hazmat records (all levels)
sm_recs	-sum of the CR safety management records (all levels)
vh_recs	-sum of the CR vehicle records (all levels)
dri	-driver review indicator
hmri	-hazmat review indicator
smri	-safety management review indicator
vri	-vehicle review indicator
rvcnfvri	- vehicle review confidence level

rvcnfdri	- driver review confidence level
lstrevdt	-date of the last CR
clean_cr	-1 if there are no CR violations; 0 if there are one or more CR violations
agewt	-age weight applied to recordable crashes in CR
rpa	-recordable preventable accidents (no longer used)
recordb	-recordable crashes in CR
vmt	-vehicle miles travelled
rar	-recordable accident rate
ra_grup	-recordable accident peer group
tmprai	-temporary review accident indicator
rai	-review accident indicator
pu_cur	-power units only used in computing the aii. when it exists the value is identical to pu value above
fatali_1	-number of fatalities in state reported crashes
injuri_1	-number of injuries in state reported crashes
hm_incd1	-number of hazmat release events
tctwa	-total NGA accident measure.
tctwa_nu	-total NGA accident measure since the last CR
n_accinu	-number of state reported crashes since the last CR
lst_crsh	-age of the newest state reported crash (months)
n_acci	-number of state reported crashes
acc_grup	-state reported crash peer group
no_motor	-1 if there are missing or zero power unit values
aim	-accident involvement measure
rankaii	-rank value of the aii within each peer group
tmpaii	-temporary accident involvement indicator
aii	-accident involvement indicator
drvinter	-number of interstate drivers
drvintra	-number of intrastate drivers
drvtleas	-number of leased drivers
mvt	-sum of the time-weighted number of moving violations
n_mv	-number of moving violations in the last 30 months
mv_grp	-moving violation peer group
tt_drive	-total number of drivers
mvm	-moving violations measure
tmpmvi	-temporary moving violation indicator
rk_mvi	-rank value of the mvi within each peer group
mvi	-moving violation indicator
maxdr_i	-max value between the dri and the dii
drscore	-driver score: an intermediate computation of the driver SEA
drsea	-driver SEA
acsea	-accident SEA
vhscore	-vehicle score: an intermediate computation of the vehicle SEA
vhsea	-vehicle SEA
smscore	-safety mngt score: an intermediate computation of the safety mngt SEA
smsea	- safety management SEA
seatot	-sum of weighted deficient SEAs
seacount	-number of deficient SEAs
seascore	-sum of weighted deficient SEAs when there are 2 or more deficient SEAs
sea_cat	-SEA category
ssscore	-SafeStat Score
rankmeas	-a temporary intermediate score used to compute SafeStat rank
ssrank	-national rank based on the SafeStat Score
ssrankst	-state rank based on the SafeStat Score

The next two fields are discussed in the section above on: Time and Severity Weighted Post-Simulation Crashes

num\_acc - the number of state-reported crashes during the post-simulation period  
new\_acc - the number of weighted state-reported crashes during the post-simulation period.

Data Dictionary for fields in **ss\_eff02.sav**

- SafeStat Results produced in the Updated 2003 Effectiveness Study (118,907 carriers)

-

These carriers are a subset of the SafeStat\_Simulation2.sav. The carriers in this file have a value of 1,2 or 3 in the 'Type' field, have a nonzero power unit value, and participated in the full 18 month post-simulation period.

dot#	-U.S. DOT Number of the carrier
pu_sim	-Number of Power Units based on the September 2002 MCMIS snapshot; this was used in the SafeStat simulation.
state_co	-State of Residence
carrier_	-Carriers Name
types	- 1 indicates the carrier has a SEA Category value of A or B 2 indicates the carrier has a SEA Category value of C 3 indicates the carrier has sufficient data but 'not selected' as having a safety concern
xssscore	- SafeStat Score
xacsea	- Accident SEA
xdrsea	- Driver SEA
xvhsea	- Vehicle SEA
xsmsea	- Safety Management SEA

The names of the next 14 fields end in 'grp' and each have values from 1 to 20. Each is derived from a SafeStat SEA or indicator. The value is computed as follows:

$$xxseagr = 1 + \text{trunc}(xxsea / 5)$$

If xxsea is 100, then xxseagr is set to 20.

Acseagr	- derived from Accident SEA
drseagr	- derived from Driver SEA
vhseagr	- derived from Vehicle SEA
smseagr	- derived from Safety Management SEA
aii_grp	- derived from All
rai_grp	- derived from RAI
dii_grp	- derived from DII
dri_grp	- derived from DRI
mvi_grp	- derived from MVI
vii_grp	- derived from VII
vri_grp	- derived from VRI
ehi_grp	- derived from EHI
hmri_grp	- derived from HMRI
smri_grp	- derived from SMRI
xrpar	- recordable accident measure (related to RAI)
xvmt	- vehicle miles traveled in the year previous to the CR (related to RAI)
xn_acci	- number of state-reported crashes (related to All)
xno_enf	- total number of enforcement
xtt_drin	- total number of driver inspections in the previous 30 months
xtt_vhin	- total number of vehicle inspections in the previous 30 months
xtt_do	- total number of driver OOS inspections in the previous 30 months
xtt_vo	- total number of vehicle OOS inspections in the previous 30 months
xrev_age	- age of the CR (in months)
xtt_hmo	- total number of hazmat OOS inspections in the previous 30 months
xrpai	- Recordable Accident Indicator (RAI)
xaii	- Accident Involvement Indicator (All)

xdii	- Driver Inspection Indicator (DII)
xdri	- Driver Review Indicator (DRI)
xvri	- Vehicle Review Indicator (VRI)
xvii	- Vehicle Inspection Indicator (VII)
xmvi	- Moving Violation Indicator (MVI)
xhmri	- Hazmat Review Indicator (HMRI)
xehi	- Enforcement History Indicator (EHI)
xsmri	- Safety Management Review Indicator (SMRI)
idcutdat	- Date of the SafeStat snapshot
lstrevdt	- Last CR date
lstenfdt	- Last Enforcement date
overall_	- Overall CR rating
sea_cat	- SEA Category
chk_a	- 1 if a CR was performed in the last 18 months, else 0
chk_b	- 1 if there are 3 or more driver inspections in the last 30 months, else 0
chk_c	- 1 if there are 3 or more vehicle inspections in the last 30 months, else 0
chk_d	- 1 if there are 3 or more HM OOS inspections in the last 30 months, else 0
chk_e	- 1 if there was an enforcement in the last 6 years, else 0
chk_f	- 1 if there was 2 or more state-reported crashes in the last 30 months, else 0
acsea_h	- if accident sea is missing then 0, if accident sea is NOT missing then 1
drsea_h	- if driver sea is missing then 0, if driver sea is NOT missing then 1
vhsea_h	- if vehicle sea is missing then 0, if vehicle sea is NOT missing then 1
smsea_h	- if safety mgnt. sea is missing then 0, if safety mgmt. sea is NOT missing then 1
aii_h	- if All is missing then 0, if All is NOT missing then 1
rai_h	- if RAI is missing then 0, if RAI is NOT missing then 1
vii_h	- if VII is missing then 0, if VII is NOT missing then 1
vri_h	- if VRI is missing then 0, if 1 then VRI is NOT missing then 1
dii_h	- if DII is missing then 0, if DII is NOT missing then 1
dri_h	- if DRI is missing then 0, if DRI is NOT missing then 1
mvi_h	- if MVI is missing then 0, if MVI is NOT missing then 1
smri_h	- if SMRI is missing then 0, if 1 then SMRI is NOT missing then 1
hmri_h	- if HMRI is missing then 0, if 1 then HMRI is NOT missing then 1
ehi_h	- if EHI is missing then 0, if 1 then EHI is NOT missing then 1
pu	- power units associated with the post simulation period (Sep-02 snapshot)
num_acc	- number of state reported crashes found in the post simulation period
new_acc	- number of weighted state reported crashes found in the post simulation period
fatali_1	- number of fatalities associated with the post simulation period
injuri_1	- number of injuries associated with the post simulation period
thrsh	- upper bound threshold value as a function of power unit value. A carrier is identified as an outlier if the number of crashes in the post-simulation period exceeds thrsh.
lowbound	- lowbound = 1 if the carrier is identified as an outlier due to its low crash rate.
cvis	- 'P' indicates that the carrier is an outlier and therefore <u>not</u> included in the evaluation; 'N' indicates that the carrier is NOT an outlier and therefore <u>included</u> in the evaluation
xdr	- if the carrier has a Driver SEA of 75 or greater then xdr = 1; else xdr = 0.
xvh	- if the carrier has a Vehicle SEA of 75 or greater then xvh = 1; else xvh = 0.
xsm	- if the carrier has a Safety Mgmt. SEA of 75 or greater then xsm = 1; else xsm = 0.
xac	- if the carrier has an Accident SEA of 75 or greater then xac = 1; else xac = 0.
xaii_f	- if the carrier has an All of 75 or greater then xaii_f = 1; else xaii_f = 0.
xrpai_f	- if the carrier has an RAI of 75 or greater then xrpai_f = 1; else xrpai_f = 0.
xdii_f	- if the carrier has an DII of 75 or greater then xdii_f = 1; else xdii_f = 0.
xdri_f	- if the carrier has an DRI of 75 or greater then xdri_f = 1; else xdri_f = 0.

xvii_f	- if the carrier has an DRI of 75 or greater then xvii_f = 1; else xvii_f = 0.
xvri_f	- if the carrier has an VRI of 75 or greater then xvri_f = 1; else xvri_f = 0.
xmvi_f	- if the carrier has an MVI of 75 or greater then xmvi_f = 1; else xmvi_f = 0.
xhmri_f	- if the carrier has an HMRI of 75 or greater then xhmri_f = 1; else xhmri_f = 0.
xehi_f	- if the carrier has an EHI of 75 or greater then xehi_f = 1; else xehi_f = 0.
xsmri_f	- if the carrier has an SMRI of 75 or greater then xsmri_f = 1; else xsmri_f = 0.
acc_rate	- the number of weighted crashes found in the post simulation period divided by the power units

## Appendix B:

### Crash Rate Score Confidence Intervals

Suppose  $L$  and  $U$  denote lower and upper  $1-\alpha$  confidence bounds for a Poisson mean  $\lambda$ , possibly at specified levels of independent variables in a Poisson regression. Such confidence bounds can be computed easily with software such as the SAS Genmod procedure. In the CR estimation problem,  $\lambda$  will denote a mean *per PU*, and confidence intervals considered here will be ranges for single PUs (not totals over multiple PUs for carriers).

Consider the problem of predicting a new Poisson random variable  $X$  (for a single PU) with the same mean  $\lambda$ . For a specified probability  $\alpha^*$ , we would like a lower prediction bound  $L^*$  such that  $P(X < L^*) = \alpha^*$ . An analogous derivation leads to an upper prediction bound  $U^*$ . Observe that

$$\begin{aligned} P(X < L^*) &= P(X < L^* \mid \lambda > L) P(\lambda > L) + P(X < L^* \mid \lambda \leq L) P(\lambda \leq L) \\ &\leq P(X < L^* \mid \lambda = L) P(\lambda > L) + P(\lambda \leq L) = F(L^*; L) (1-\alpha) + \alpha, \end{aligned}$$

where  $F(x; \eta)$  denotes the cumulative distribution function at  $x$  of a Poisson random variable with mean  $\eta$ . Let  $q = (\alpha^* - \alpha) / (1-\alpha)$ , and take  $L^*$  as the  $q^{\text{th}}$  quantile of  $F(\cdot; L)$ . Then  $P(X < L^*) \leq q(1-\alpha) + \alpha = ((\alpha^* - \alpha) / (1-\alpha)) (1-\alpha) + \alpha = (\alpha^* - \alpha) + \alpha = \alpha^*$ .

$L^*$  can be solved for (for specified  $\alpha^*$  and thus  $q$ ) using the Poisson distribution with mean  $L$ . This is how the confidence intervals in Table 7 were derived. Similar confidence intervals can also be derived for distributions other than the Poisson, for example the negative binomial, although additional approximations (or more difficult mathematics) are generally required. For example, two estimated parameters must be dealt with for the negative binomial distribution, as opposed to a single parameter for the Poisson distribution.