FINAL REPORT

STATISTICAL ESTIMATES OF TOTAL REPORTED HAZARDOUS MATERIAL INCIDENTS COSTS

to

U.S. Department of Transportation
Pipeline and Hazardous Materials Safety Administration
Office of Hazardous Materials Safety

from

Battelle
The Business of Innovation

Battelle Columbus Operations
505 King Avenue
Columbus, Ohio 43201-2693

September 2005

TABLE OF CONTENTS

PAGE

TABLE OF CONTENTS (CONTINUED)

PAGE

List of Tables

List of Tables in Appendices

List of Figures

List of Figures in Appendices

EXECUTIVE SUMMARY

The Pipeline and Hazardous Materials Safety Administration's (PHMSA[1]) Office of Hazardous Materials Safety (OHMS) maintains the Hazardous Materials Incident Reporting System (HMIRS), a database of information regarding reported incidents involving hazardous materials. Lack of guidance and definitions to be used in providing data for the financial consequences and under-reporting by the regulated community seriously compromise the credibility of the cost fields of the HMIRS database. OHMS has commissioned a study to examine the benefits of creating a statistical model to provide estimates of the total annual financial consequences associated with hazardous material incidents. By obtaining more accurate cost information for a sample of reported hazardous material incidents, creating statistical models to predict the more accurate costs, and applying the model to all reported hazardous material incidents with a single year, an improved estimate of the total costs of those incidents can be obtained.

A stratified sample of 500 reported hazardous material incidents that occurred between November 1, 2002, and October 31, 2003, was selected. This sample included 101 high-cost incidents, 248 highway incidents, 130 rail incidents, 14 air incidents, and 7 water incidents. Contacts listed in the HMIRS database for each reported incident were called and asked to provide more detailed cost information regarding the incident for five categories. Complete responses were obtained for 260 of these incidents, and partial cost data (not all of the categories) were obtained for about 150 of the remaining sampled incidents. The more accurate incident costs, along with a set of potential predictor variables obtained from the HMIRS database, were used to obtain four statistical estimates of total reported incident costs over the specified period. These estimates included stratified sampling estimates and stratified regression estimates using total incident cost and using costs associated with the five categories.

Among the four estimation methods, the stratified regression estimates based on cost components are recommended for use in estimating the total reported incident costs. This method is recommended because it makes the maximum use of the limited data that were available, and it produces the estimate with the smallest variability. The stratified estimate based on total cost, while providing a reasonable estimate, did not make use of any relationships between the costs and the predictor variables, nor did it use all the information available. The stratified regression estimate also was reasonably consistent with the other estimates, but it also did not make full use of the data available. The stratified sampling estimate based on component costs used all of the cost information available, but it did not use any predictor variables to reduce the variability. In addition, a few discrepancies between the sample and the population resulted in a much larger estimate than in the other three methods.

Table 1 shows the estimated incident costs for each of the five strata as well as the total estimated costs. Each cost is shown in comparison to the reported cost obtained from the HMIRS database. Table 1 indicates that the total reported incident costs are estimated to be $77.7 million, with a 95 percent confidence interval of $66.6 million to $88.8 million. This compares to HMIRS-reported costs of $49.8 million.

[1] This study was conducted for the predecessor agency to PHMSA, the Research and Special Programs Administration (RSPA). This document was finalized after PHMSA was created and, for consistency, PHMSA is used throughout the document to refer to the Administration containing the Office of Hazardous Materials Safety.

Table 1. Comparison of Estimated Total Reported Incident Costs Versus HMIRS-Reported Costs, by Stratum

Stratum	HMIRS-Reported Total Cost ($)	Estimated Total Cost ($)	Lower Confidence Bound ($)	Upper Confidence Bound ($)
High-cost	32,788,896	39,085,065	33,491,169	44,678,961
Highway	15,999,911	36,387,798	26,827,206	45,948,390
Rail	825,193	1,991,278	1,356,154	2,626,402
Air	75,723	49,632	≤ 0	112,921
Water	131,469	173,173	≤ 0	476,183
Total	49,821,192	77,686,946	66,587,586	88,786,306

Figure 1 illustrates the difference in the revised estimates as compared to the HMIRS-reported cost. The bars in Figure 1 show the HMIRS-reported costs as a percentage of the estimated cost for each stratum. Based on the results of this study, the HMIRS total cost over all strata is 64 percent of the estimated total cost. Air and water results are based on too few observations to provide accurate predictions of actual costs.

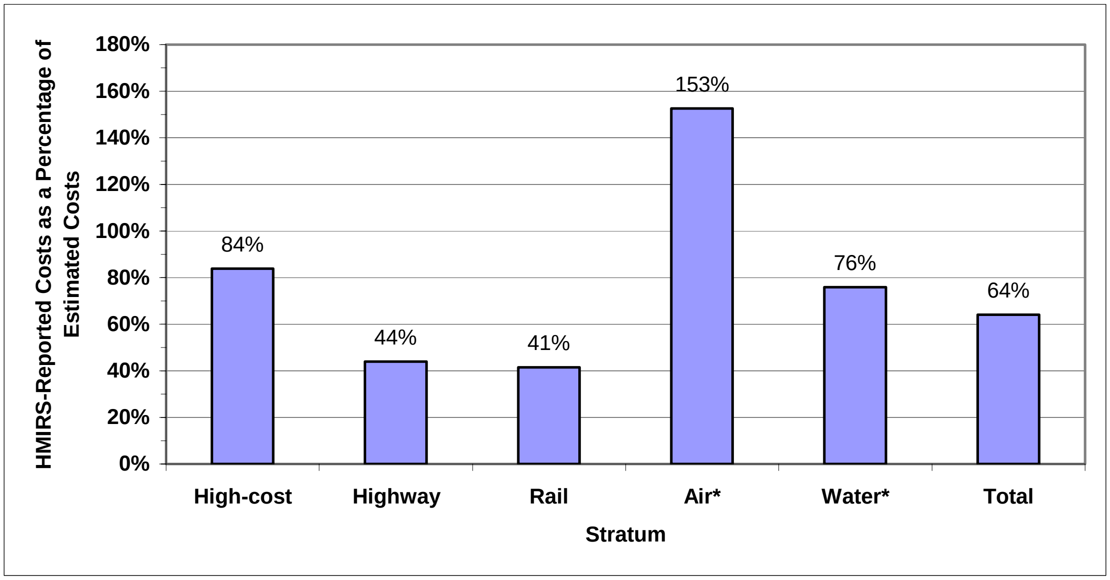


Figure 1. HMIRS-Reported Costs as Percentages of Estimated Cost

A subsequent analysis of serious incidents, as defined by PHMSA, showed that the results for serious incidents are consistent with those for incidents in the high-cost stratum. This analysis is presented in Appendix B.

1.0 INTRODUCTION

The Pipeline and Hazardous Materials Safety Administration's (PHMSA) Office of Hazardous Materials Safety (OHMS) maintains the Hazardous Materials Incident Reporting System (HMIRS), a database of information regarding reported incidents involving hazardous materials. These data are reported on Form DOT 5800.1. The current version of the form contains five financial consequence categories for incidents: product loss, carrier damage, public/private property damage, decontamination/cleanup costs, and "other" costs. When added together, these consequence costs sum to a total incident cost. Lack of guidance and definitions to be used in providing data for these categories has led to inconsistent reporting in the past. This, in addition to under-reporting by the regulated community, seriously compromises the credibility of the financial consequence fields of the HMIRS database. OHMS is currently reviewing alternatives to estimate the financial consequences associated with hazardous materials incidents.

Previously, a pilot study was performed to demonstrate the feasibility of collecting more accurate cost data from representatives of companies that were involved in such incidents. Subsequently, OHMS began a research program to establish the appropriate scope of future data collection activities such that statistically valid cost estimates could be developed. Specifically, this research program examined the relationship between sample sizes and the precision of the results that might be expected, to assist OHMS in determining the number of incidents for which more accurate cost information should be obtained. Based on analysis of ten years of historical data in the HMIRS, a study to collect more accurate cost information relating to hazardous material incidents was designed. This study consisted of calling the contacts from the carriers that were involved in the incidents and asking them to review their records to determine whether the cost data originally supplied to HMIRS could be revised as a result of more accurate and complete information being available to the organization.

PHMSA initiated the proposed study and its results are presented in this report. The primary objective of the study was to create a statistical model that can provide estimates of the total annual financial consequences associated with reported hazardous material incidents (i.e., costs associated with incidents not reported to the HMIRS database are not included in the total). This objective was achieved by obtaining more accurate cost information for a sample of hazardous material incidents, creating statistical models to predict the costs, and applying the model to all hazardous material incidents with a single year to estimate the total costs of those incidents. By using a statistical model, standard errors of the cost estimates were determined that provided information on the precision of the cost estimates.

This report contains a description of the sampling design used to select incidents for which more accurate cost information was obtained, a description of the statistical methods used to obtain revised cost estimates and the precision of those estimates, results of the application of the statistical methods to the hazardous material incidents data, a discussion of the results, and a set of conclusions and recommendations.

2.0 DATA COLLECTION

2.1 Sample Selection

A previous study examined the relationship between sample sizes and the precision of the results that might be expected to assist OHMS in determining the number of incidents for which more accurate cost information should be obtained. While that study showed that the precision of the statistical incident cost estimates are improved when the sample size is increased, it also showed that there was a significant level of disagreement between observed costs and predicted costs. Furthermore, both the estimates and their precision could be improved by employing a stratified sampling approach, stratifying the incidents by the mode of transportation, and considering high-cost incidents as a separate stratum. The improvement in precision is gained by eliminating the variability between modes from the uncertainty in the results.

The current study was designed to exploit the gain in precision available by stratifying the sample by mode of transportation. Based on the results of the previous study, a target sample size of 400 hazardous material incidents, distributed among five strata, was selected. Four of the strata were based on four modes of transportation: air, highway, rail, and water. The remaining stratum was defined to be "high-cost" incidents. These were separated from other incidents to improve the precision of the final estimate but also because these incidents have the greatest impact on the total financial consequences and, thus, are of greatest interest. A high-cost incident was defined to be any incident for which at least one of the following was true:

- there was at least one death;
- there was at least one major injury;
- there was a radioactive release;
- there were at least 100 people evacuated;
- a major road was closed; or
- the reported costs in HMIRS exceeded $100,000.

The targeted sample was to include all high-cost incidents (which were estimated to be about 100 per year), all water incidents (approximately 10 per year), and 300 additional incidents from strata defined by the air, highway, and rail modes of transportation. The specific numbers of incidents in each of those three strata are chosen in proportion to total reported costs among all incidents within the mode in order to provide the most representative data for assessing incident costs. The sample was to be drawn from the population of all incidents reported in the HMIRS database with incident dates between November 1, 2002, and October 31, 2003. These dates were chosen to balance the timeliness of incident data with the need for sufficient time since the incidents occurred to allow additional costs that were unknown or only estimated at the time of the report submission to be collected or updated. The actual number of incidents included in the sample was greater than the targeted 400 to account for nonparticipation by the reporting organizations.

At the outset of the project, the HMIRS database was not expected to have entries for all incidents occurring within the selected dates because of the expected lag time required to submit and incorporate the data into HMIRS. To make the best use of time, incidents were selected in two draws so that the interviews could begin before the HMIRS database entries for the sampling period were complete. The first draw was performed in early November 2003, when the HMIRS contained data for most incidents occurring before August 2003, and the second was performed in early February 2004. The following protocol was used to select the first sample:

- All HMIRS records for incidents occurring between November 1, 2002 and October 31, 2003 were downloaded into an Excel spreadsheet. The spreadsheet data were then read into a SAS program for processing. The Excel spreadsheet was saved for use with the second sample.
- There were cases where there was more than one record for an incident (as identified by RPTNO in the HMIRS database). For sampling purposes, a single record per incident was required, so separate records for the same incident were combined into a single record. Only those variables that were needed to identify the incident, its mode of transportation, and whether the incident qualified as a high-cost incident where retained in the reduced data set. Variables that were counts or costs (numbers of deaths, major injuries, evacuations, reported costs) were summed over all records within an incident, and indicator variables (radioactive release, highway closure) were set to indicate whether any of the records had positive results. This data set comprised the sampling frame for the first sample.
- Each incident was evaluated to determine whether it qualified as a high-cost incident. Those that were so identified were removed and placed in a separate data set, which was designated as “selected high-cost incidents.”
- All remaining incidents that had “Water” as the mode of transportation were removed and placed in a separate data set, which was designated as “selected water incidents.”
- The remaining data were divided into three separate data sets for air, highway, and rail incidents.
- A random sample of eight air incidents was selected by assigning each incident a uniform random number between 0 and 1 and choosing the eight incidents with the smallest random number. Random numbers were selected using the RANUNI function in SAS with a starting seed determined by the time that the program was run. The selected incidents were placed in a separate data set, which was designated as “selected air incidents.”
- A random sample of 160 highway incidents was selected using the same process used with the air sample. They were placed in a separate data set designated “selected highway incidents.”
- A random sample of 80 rail incidents was selected in a similar manner as air and highway incidents, with the selected incidents placed is a separate data set designated “selected rail incidents.”
- The five data sets containing the sampled incidents were combined into a single data set.

- For each sampled incident, all records in the original downloaded database where placed in a new database that defined the sampled incidents. This database was provided to the interviewers.

The sample-selection protocol for the second sample was as follows:

- All HMIRS records for incidents occurring between November 1, 2002 and October 31, 2003 were downloaded into an Excel spreadsheet. The spreadsheet data were then read into a SAS program for processing.
- Incident records were reduced to one per incident as described in the protocol for the first sample.
- The resulting list of incidents was compared to the sampling frame for the first sample, and all incidents that appeared in the first sampling frame were removed, leaving only incidents that were new to the database since the selection of the first sample. This list comprised the sampling frame for the second sample.
- All high-cost and water incidents in the new sampling frame were selected. Random samples of 6 air incidents, 90 highway incidents, and 50 rail incidents were selected in a similar manner to the first sample.
- For each sampled incident, all records from the original database were downloaded and added to the sample database created for the first sample. The revised database was provided to the interviewers.

One shortcoming of this sampling method is that there were unequal probabilities of selection for all incidents within each stratum. The number of incidents available during the first sample was significantly larger than the number available during the second sample. As a result, the probability of selection was higher for incidents in the second sample. However, the difference in selection probabilities between the first and second sampling periods was generally small enough to be considered unimportant, and no weighting adjustments were performed.

The final sample that was selected, after corrections and deletions of incidents that were not actually hazardous-material incidents, included:

- all water incidents (of which there were 7);
- all high-cost incidents (101);
- 14 air incidents;
- 248 highway incidents; and
- 130 rail incidents.

2.2 Data Collection

Between December 2003 and July 2004, individuals listed in the HMIRS database for each of the selected incidents were contacted by telephone. The data collection team used a custom database management system to track all contacts, correspondence, and activity related to each incident.

In many cases, the original contact was not available, the contact information originally submitted to DOT was inaccurate, or the company could not be located. For each of these incidents, the team member assigned to that incident used a combination of sources to identify the appropriate company contact information and locate the person responsible for hazardous materials incident reporting. In some cases, a search in the HMIRS for other reports from the same company yielded alternative individuals to contact and, in other cases, Internet searches could identify main numbers for reporting entities.

Multiple calls were usually required to obtain all cost information; including calls to other carrier personnel, the shipper, cleanup contractors, response companies, and police and fire departments. Initial telephone contact with each entity was followed by an e-mail containing a more detailed itemization of the requested information, customized for each incident and for the type of entity being contacted. Data collection continued until records were complete, the contacts ceased to cooperate, or third-party sources of information were exhausted. Where a contact was unable to provide direct cost estimates for a particular item, the data collection team member would try to piece together component costs to estimate an approximate cost. For example, if a fire department that responded to an incident did not seek reimbursement for their response costs or track them in any way, it was usually possible to speak to the on-scene incident commander. He or she would be able to recall how many people were on scene, the length of time they were deployed, their approximate hourly wages, any equipment that was expended during the incident response, and other relevant information.

Incident cost information was obtained for the five cost categories on the new incident report form required for incidents occurring after January 1, 2005 (material loss, carrier damage, property damage, response cost, cleanup cost), using several subcategories within each category to assist the incident manager in identifying applicable costs. A list of the subcategories can be found in Appendix C, which contains the supporting documentation provided to the entities from which we requested data.

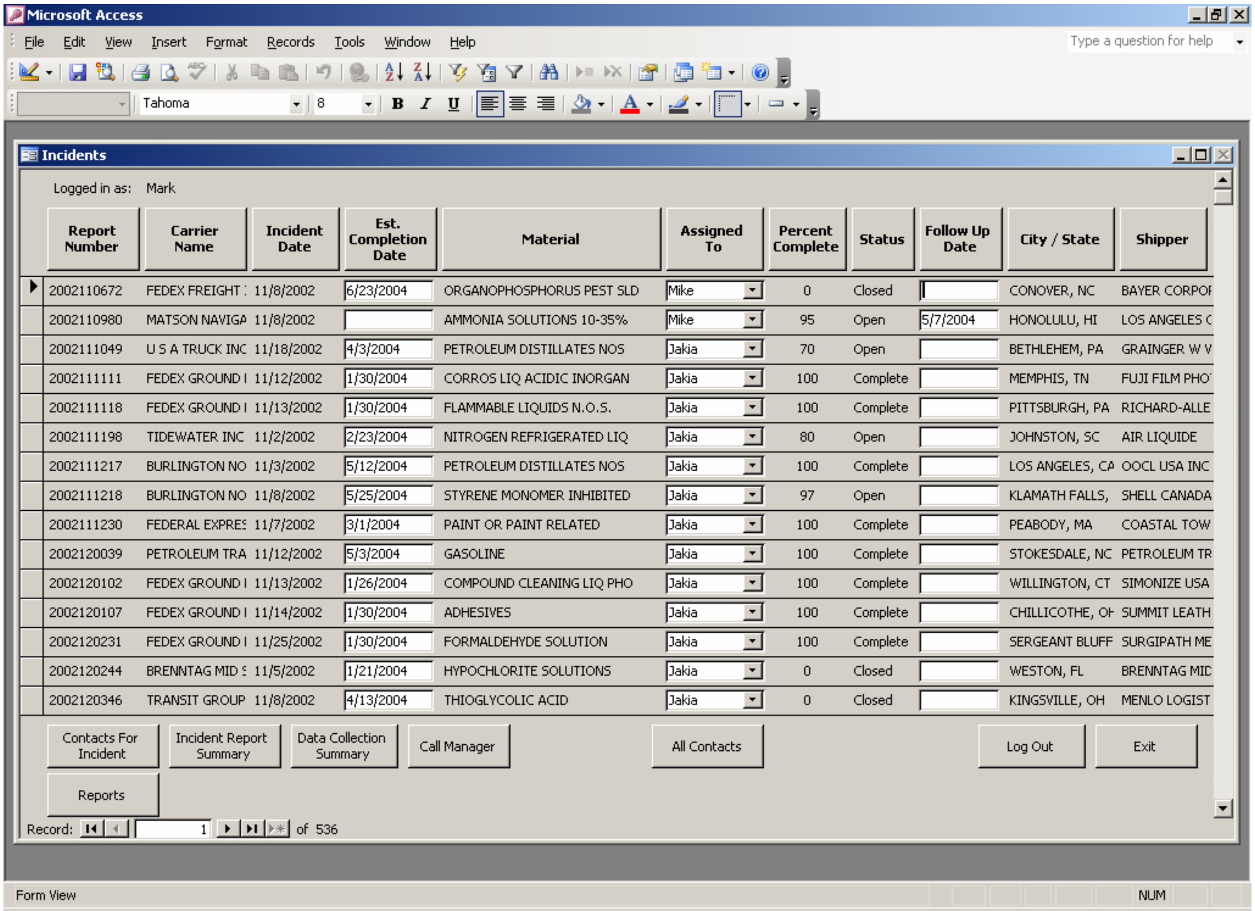
Throughout the data collection process, the data collection team continually estimated the percent completion for each of the five cost categories. In other words, based on their understanding of the costs that were incurred or likely to have been incurred, they would estimate the percentage of those costs that they had thus far been able to quantify from all sources. This provided the basis for subsequent filtering of incidents for the statistical modeling.

2.3 Databases

The 2002 and 2003 HMIRS databases were downloaded from the HMIRS website in November 2003. Database records (both Material and Container data) for those incidents selected in the sampling process were imported into a Microsoft Access database that was designed to manage the calling process. The 2002 and 2003 HMIRS databases from February 2004 were also downloaded and used to choose a supplemental sample of incidents not included in the original sampling frame. The complete records (Material and Container) were imported into the Call Manager database to form the sampling database. The Call Manager database enabled the interviewers to perform a variety of tasks, including:

- view the complete Incident Report as filed for each incident;
- indicate the status (e.g., data provided, busy signal, left message) of each call;
- enter cost data obtained during the call;
- view cost data collected during previous calls;
- add additional contacts for an incident;
- track the completeness status for each incident; and
- view summary reports for each incident and for the entire database.

The Material and Container data for a particular incident report were linked via the “Report Number” field in the database. Other linking variables were created as necessary to link the cost data with their respective incidents. Figure 2 illustrates the main screen from the Call Manager database.



Logged in as: Mark

	Report Number	Carrier Name	Incident Date	Est. Completion Date	Material	Assigned To	Percent Complete	Status	Follow Up Date	City / State	Shipper
▶	2002110672	FEDEX FREIGHT :	11/8/2002	6/23/2004	ORGANOPHOSPHORUS PEST SLD	Mike	0	Closed		CONOVER, NC	BAYER CORPOI
	2002110980	MATSON NAVIGA	11/8/2002		AMMONIA SOLUTIONS 10-35%	Mike	95	Open	5/7/2004	HONOLULU, HI	LOS ANGELES C
	2002111049	U S A TRUCK INC	11/18/2002	4/3/2004	PETROLEUM DISTILLATES NOS	Jakia	70	Open		BETHLEHEM, PA	GRAINGER W V
	2002111111	FEDEX GROUND I	11/12/2002	1/30/2004	CORROS LIQ ACIDIC INORGAN	Jakia	100	Complete		MEMPHIS, TN	FUJI FILM PHO
	2002111118	FEDEX GROUND I	11/13/2002	1/30/2004	FLAMMABLE LIQUIDS N.O.S.	Jakia	100	Complete		PITTSBURGH, PA	RICHARD-ALLE
	2002111198	TIDEWATER INC	11/2/2002	2/23/2004	NITROGEN REFRIGERATED LIQ	Jakia	80	Open		JOHNSTON, SC	AIR LIQUIDE
	2002111217	BURLINGTON NO	11/3/2002	5/12/2004	PETROLEUM DISTILLATES NOS	Jakia	100	Complete		LOS ANGELES, CA	OOCL USA INC
	2002111218	BURLINGTON NO	11/8/2002	5/25/2004	STYRENE MONOMER INHIBITED	Jakia	97	Open		KLAMATH FALLS,	SHELL CANADA
	2002111230	FEDERAL EXPRES	11/7/2002	3/1/2004	PAINT OR PAINT RELATED	Jakia	100	Complete		PEABODY, MA	COASTAL TOW
	2002120039	PETROLEUM TRA	11/12/2002	5/3/2004	GASOLINE	Jakia	100	Complete		STOKESDALE, NC	PETROLEUM TR
	2002120102	FEDEX GROUND I	11/13/2002	1/26/2004	COMPOUND CLEANING LIQ PHO	Jakia	100	Complete		WILLINGTON, CT	SIMONIZE USA
	2002120107	FEDEX GROUND I	11/14/2002	1/30/2004	ADHESIVES	Jakia	100	Complete		CHILLICOTHE, OH	SUMMIT LEATH
	2002120231	FEDEX GROUND I	11/25/2002	1/30/2004	FORMALDEHYDE SOLUTION	Jakia	100	Complete		SERGEANT BLUFF	SURGIPATH ME
	2002120244	BRENNTAG MID S	11/5/2002	1/21/2004	HYPOCHLORITE SOLUTIONS	Jakia	0	Closed		WESTON, FL	BRENNTAG MID
	2002120346	TRANSIT GROUP	11/8/2002	4/13/2004	THIOGLYCOLIC ACID	Jakia	0	Closed		KINGSVILLE, OH	MENLO LOGIST

Record: 1 of 536

Figure 2. Main Screen from Call Manager Database

Once the data collection period had ended, database queries were run to collapse the data into a format suitable for statistical analyses. The process of collapsing the data included the following steps:

- categorical variables were converted into multiple indicator variables (one indicator per component value – for example, the HMIRS variable indicating type of community was replaced with three indicator variables, one each for urban, suburban, and rural);

- multiple Container records for a particular Material record were combined into one record by summing or averaging numeric variables (e.g., number evacuated) as appropriate and setting indicator variables (e.g., runoff into sewer?) to "true" if any of their respective values on the individual records were set to "true";

- cost data obtained during the interviews were merged with their corresponding Material records;

- multiple records with the same "Report Number" (both Material and Container records) were merged, yielding one record per incident.

The 2002 and 2003 HMIRS databases were again downloaded from the HMIRS web site in August 2004 to capture any updates or additions. Database records (both Material and Container data) for all incidents occurring between November 1, 2002, and October 31, 2003, were imported into separate tables in the Microsoft Access database. Similar collapsing of data as described above for the Call Manager database (with the exception of the inclusion of interview-obtained cost data) was performed on the complete HMIRS data.

3.0 STATISTICAL ANALYSIS METHODS

The objective of this study was to obtain estimates of the total annual financial consequences of reported hazardous material incidents over a one-year period based on revised cost information obtained from a sample of incidents. A statistical analysis of the sample of revised costs, along with other auxiliary information, was applied to a set of incident data for an entire year to estimate the true costs of the reported hazardous materials incidents for that year's worth of data. The source of the revised costs was contacts from organizations associated with reported incidents. The population from which the sample was drawn was the HMIRS database. That database also was the source of the auxiliary information and the source of the year-long data that were used to form the basis of the total incident cost estimate.

All statistical analyses were performed by writing SAS® programs that directly accessed the Microsoft Access study database. As described below, these SAS® programs also filtered out selected records prior to performing the analyses.

- Only those records for which the interviewers provided an estimated completion percentage of at least 90 percent and obtained cost data were used. In three cases, the interviewers recorded 0 percent completion but indicated in the comments that the only missing costs were for police and fire department responses; these cases also were included.

- Four sampled events were excluded, two because they were duplicates of other sampled records, and the other two because they were not hazardous material incidents that should have been included in the HMIRS database.

3.1 Stratified Sampling Estimate

Stratified sampling, which is the method that was used to draw the sample of hazardous material incidents, provides a method for obtaining estimates of population totals that is useful when a population is divided into several groups whose members are similar but which may differ greatly from members of the other groups. This method was used to obtain separate total cost estimates for each stratum separately and then combine them into a single overall estimate. The benefit of this method is that the variability of the total cost estimate includes only within-group variability and not between-stratum variability, which is often a large component of the overall variability in incident costs.

For this study, stratification was based on the mode of transportation, with potential high-cost incidents separated out into their own stratum. Specifically, the five strata used in this study were:

- potential high-cost incidents;
- highway incidents (other than those included in the potential high-cost incident stratum);
- rail incidents (other than those included in the potential high-cost incident stratum);

- air incidents (other than those included in the potential high-cost incident stratum); and
- water incidents (other than those included in the potential high-cost incident stratum).

In two of the strata (high-cost, water), all available incidents were selected; for the remaining three strata (highway, rail, air), separate samples were selected, with the numbers of samples proportional to the total reported cost in each stratum (estimated over a 10-year period).

Traditional stratified sampling methods can be used to obtain an estimate of the total costs for reported hazardous material incidents that is based only on the revised cost information obtained from the sampled incidents. Mathematically, this estimate of total costs, which is denoted by $\hat{\tau}_{st}$, can be written as

$$\hat{\tau}_{st} = \sum_{k=1}^{K} \hat{\tau}_k = \sum_{k=1}^{K} N_k \bar{y}_k \,, \qquad (1)$$

where K is the number of strata, N_k is the number of incidents in stratum k, and $\bar{y}_k$ is the average cost among the sampled incidents in stratum k. The precision of this estimate of total reported incident costs can be expressed in terms of the variance of the estimate, which can be written as

$$Var(\hat{\tau}_{st}) = \sum_{k=1}^{K} Var(\hat{\tau}_k) = \sum_{k=1}^{K} N_k (N_k - n_k) \frac{s_k^2}{n_k} \,, \qquad (2)$$

where s_k^2 is the estimated variance of the observed costs in the kth stratum, and n_k is the number of incidents sampled in stratum k.

Using the estimated total costs and corresponding variance, a confidence interval can be obtained for the true total reported incident costs. This confidence interval can be written as

$$\hat{\tau}_{st} \pm z_{(1-\alpha/2)} \sqrt{Var(\hat{\tau}_{st})} \,, \qquad (3)$$

where $z_{(1-\alpha/2)}$ is the 100 (1-α/2) percentile of the normal distribution and 100 (1-α/2) is the confidence level. For a 95 percent confidence interval, which was used in this study, the value of $z_{(0.95)}$ is 1.96.

3.2 Regression Estimate

One of the weaknesses of the stratified sampling method is that it makes no use of any information about the incidents except for the cost. It would be expected for the cost to be a function of certain characteristics of the incident. For example, an incident in which a death occurred likely would incur a higher cost than a similar incident in which there is not a death. An improved estimate can be obtained when the auxiliary information about the incident is incorporated into the estimation method. One such method that makes use of the auxiliary information is regression estimation. In this method, the relationship between cost and the predictor (auxiliary) variables is determined for a sample of incidents, and that relationship is applied to all incidents to estimate the total cost.

Because there are often significant differences between incident costs among the various strata, it is natural to combine both the stratified and regression approaches. In the stratified regression approach, separate regression estimates are obtained for each stratum and the results are summed over all strata to produce a total reported incident cost estimate. This approach allows different predictor variables to be used for each of the strata, which may be more reasonable than requiring all strata to have the same set of predictors, because different incident characteristics may be more applicable to one stratum than another. For example, the type of highway on which an incident occurs may be important in determining the cost for a highway incident, but it will not be applicable to rail, air, or water incidents.

Mathematically, the regression estimate of the total incident costs within a stratum, which is denoted by $\hat{\tau}_{Rk}$, can be written as

$$\hat{\tau}_{Rk} = \sum_{i=1}^{N_k} \hat{y}_i = \sum_{i=1}^{N_k} \left(a + \sum_{j=1}^{J_k} b_j x_{ji} \right), \qquad (4)$$

where $\hat{y}_i$ is the predicted value of the i-th incident in the k-th stratum, J_k is the number of predictors used for the k-th stratum, x_{ji} is the value of the j-th predictor variable for the i-th incident in the k-th stratum, b_j is the linear relationship between the j-th predictor and the incident cost, and a is the intercept (cost when all predictors are equal to zero). The variance of the regression estimator can be written as

$$Var(\hat{\tau}_{Rk}) = \frac{N_k (N_k - n_k)}{n_k (n_k - J_k)} \sum_{i=1}^{N_k} (y_i - \hat{y}_i)^2 = \frac{N_k (N_k - n_k)}{n_k} MSE_k, \qquad (5)$$

where N_k is the number of incidents in the k-th stratum, n_k is the number of incidents sampled in the k-th stratum, y_i is the actual cost of the i-th incident, and MSE_k is the mean square error from the regression fit in the k^{th} stratum. The stratified regression estimate for the overall total reported incident costs is obtained by summing the regression estimates for each stratum, and its variance is the sum of the variances for each stratum. A confidence interval can be obtained using the format of Equation (3).

3.3 Cost Component Estimates

One of the difficulties with using any of the methods above, which are used to estimate the total reported incident costs directly, is that the data are not fully used. The total incident cost is defined as the sum of five cost components. Only when there are complete data for all of the components can a total incident cost be obtained. There are a number of incidents where the callers were able to obtain costs for some of the categories during the interview process but not for all of the categories. In such cases, the component costs are valid, but a total cost was not obtained, thus the data were not used to determine the final estimate.

As an alternative to "wasting" valid data, separate cost estimates could be obtained for each of the cost categories using all the available data, and the component estimates could be combined to obtain a final total incident cost estimate. This method could be used with the stratified sampling estimate, the regression estimate, or the stratified regression estimate. The total component cost (across all incidents) would be obtained using the formulas shown above, as would the variance for each cost component. The total cost over all categories would be estimated as the sum of the component costs. The variance of that estimate is a function of the variance of the individual component costs as well as the correlations between the cost categories. For instance, higher cleanup costs may be more likely to occur for incidents that also have a large material loss, resulting in a positive correlation between the two component costs. The correlations between component costs affect the overall variance of the total incident costs. Specifically, that variance would be written as

$$Var(\hat{\tau}_c) = \sum_{l=1}^{5} Var(\hat{\tau}_l) + 2\sum_{l<m} \rho_{l,m} \sqrt{Var(\hat{\tau}_l) \cdot Var(\hat{\tau}_m)}\,, \qquad (6)$$

where $\hat{\tau}_c$ is the total cost obtained by summing component totals, $\hat{\tau}_l$ and $\hat{\tau}_m$ are the costs associated with the l-th and m-th categories, and $\rho_{l,m}$ is the correlation between the l-th and m-th component. In the case of the stratified sampling approach, the correlation is estimated by the correlation between the two cost components, while with the regression approach, it is estimated by the partial correlation coefficient obtained by fitting the regression model to the components as a group.

4.0 RESULTS

4.1 Data Summary

The database on which the statistical analysis is based is comprised of two parts. The first part consists of the set of sampled reported incidents and includes revised cost estimates for five categories as well as all of the HMIRS data from the Material and Container tables for those records. Some of these records had complete revised cost information, while others did not have some or all of the revised costs. The second part consists of all Material and Container records for incidents occurring between November 1, 2002, and October 31, 2003, that were contained in the HMIRS database on the PHMSA website on August 4, 2004, when the data were downloaded for the statistical analysis.

Table 2 provides a summary of the number of records that were included in this study. This table includes a breakdown by the five sampling strata. Table 2 indicates that 52 percent of the incidents in the sample (260 of 500) had complete revised cost information. This represents 65 percent of the target number of samples. Air incidents had the largest percent of target (80 percent) for any of the individual strata, with high-cost, highway, and rail all having between 62 and 68 percent of their targets achieved. Also of note is that the number of high-cost and water incidents sampled was less than 100 percent, which was the goal of the sampling program. This can be explained by the fact that 6 months passed between the time that the last sample was selected and when the HMIRS data were obtained for the statistical analysis. Delays in reporting and/or entering the incident information into the HMIRS database for those types of incidents resulted in new incidents being added to the database after sample selection was completed.

Table 2. Summary of Data Used in Statistical Analysis

Stratum	Number in HMIRS Database	Number Sampled	Percent Sampled	Target Number of Samples	Number with Complete Cost Information	Percent of Target for Complete Incidents
High-cost	158	101	63.9	90	59	65.6
Highway	13,503	248	1.8	190	129	67.9
Rail	800	130	16.3	100	62	62.0
Air	718	14	1.9	10	8	80.0
Water	11	7	63.6	10	2	20.0
Total	15,188	500	3.3	400	260	65.0

Table 3 provides a comparison of the cost data obtained from the study and the cost information reported in the HMIRS database. This table lists the number of observations, the number of observations where the study-obtained costs were higher than the HMIRS-reported costs, the number of incidents where the study-obtained costs were lower than the HMIRS-reported costs, the maximum of the higher costs, the maximum of the lower costs, the average cost difference, and the standard deviation of the cost differences. These are presented separately for each of the five strata and across all strata. Table 3 shows that, on average, the costs obtained in the study were higher than the HMIRS-reported costs. This supports the belief that the HMIRS costs are generally under-reported.

Table 3. Summary Statistics for Differences Between Study-Obtained and HMIRS Reported Costs, by Sampling Stratum

Statistic	Stratum					
	High-cost	Highway	Rail	Air	Water	Total
Number with Complete Costs	59	129	62	8	2	260
Number with Higher Costs	46	58	43	2	2	151
Number with Lower Costs	10	13	0	0	0	23
Maximum Higher Cost ($)	845,949	15,263	59,177	350	3,575	845,949
Maximum Lower Cost ($)	1,049,062	2,173	0	0	0	1,049,062
Average Difference ($)	75,590	647	1,835	56	1,811	17,928
Std. Dev. Difference ($)	240,354	2,163	7,625	124	2,495	118,038

Four approaches were used to obtain estimates for the total reported incident costs:

- a stratified sampling approach using the total incident costs;
- a stratified regression approach using the total incident costs;
- a stratified sampling approach using component costs; and
- a stratified regression approach using component costs.

The stratified sampling approaches were used because they provided the simplest estimate based on the study design. The stratified regression approaches were used because, unlike the stratified sampling approach, the regression approaches made use of relationships between costs and various incident characteristics. The approaches using only total incident costs were used because they provided the most direct estimates of total reported incident costs. Because the estimates based on total incident costs did not make use of any partial incident cost information, the two approaches using component costs were used to include the partial data. The results obtained using these four methods are presented in the sections that follow.

4.2 Stratified Sampling Estimate for Total Incident Cost

The first method used to obtain an estimate of total reported incident costs is the stratified sampling approach. This approach determines an average incident cost for the sample results within each stratum and expands it based on the number of reported incidents in the population for each stratum to obtain an estimate of the total stratum cost. Results are summed over the strata to produce an overall estimate. The analysis is based on total cost for each incident.

Table 4 shows the results of the stratified sampling estimation. It contains counts of incidents sampled and in the HMIRS database, average incident costs for each stratum, the standard error of that estimate, a 95 percent confidence interval for the average incident cost, the estimated total cost for the stratum, its standard error, and a 95 percent confidence interval for the total stratum cost. In addition, the total cost reported in the HMIRS database is shown for reference. Note that for the overall total, no information is shown for the average incident cost. Table 4 shows that the stratified sampling estimate of total reported incident costs is around $68 million, with a 95 percent confidence interval ranging from $47 million to $90 million.

Table 4. Stratified Sampling Estimates for Average and Total Reported Incident Costs (in Dollars), by Stratum and Overall

Statistic	Stratum					
	High-cost	Highway	Rail	Air	Water	Total
Number in Stratum	158	13,503	800	718	11	15,188
Number in Sample	59	129	62	8	2	260
Average Incident Cost ($)	263,150	1,870	2,304	69	15,583	NA
Std Error of Average Incident Cost ($)	323,090	8,204	7,721	128	21,973	NA
Lower 95% Confidence Bound for Average Incident Cost ($)	≤ 0	≤ 0	≤ 0	≤ 0	≤ 0	NA
Upper 95% Confidence Bound for Average Incident Cost ($)	896,405	17,949	17,437	320	58,649	NA
Stratum Total Cost ($)	41,051,383	25,245,529	1,843,452	49,632	171,413	68,361,409
Std Error of Stratum Total Cost ($)	5,174,223	9,706,339	753,401	32,290	154,591	11,026,248
Lower 95% Confidence Bound for Stratum Total Cost ($)	30,909,907	6,221,105	366,787	≤ 0	≤ 0	46,749,963
Upper 95% Confidence Bound for Stratum Total Cost ($)	51,192,825	44,269,953	3,320,117	112,921	474,412	89,972,855
HMIRS-Reported Stratum Total Cost ($)	32,788,896	15,999,911	825,193	75,723	131,469	49,821,192

There are several things of note in Table 4. First, as expected, the average incident cost is highest for the high-cost incidents, second highest for water incidents, and lowest for air incidents. Secondly, for the individual stratum estimates of average cost, the lower bounds of the 95 percent confidence intervals are all less than zero. This is a result of the high degree of variability in the cost even within strata. The 95 percent confidence intervals for total incident cost estimates for air and water strata also include zero, which is a result of the small number of samples obtained for those two strata. The 95 percent confidence intervals for the remaining three strata, as well as the overall total do not include zero. Thirdly, with the exception of the air stratum, the estimated total incident costs exceed the total reported incident costs, which are shown in the last row. For highway and rail incidents, the relative increase is quite large.

4.3 Stratified Regression Estimate for Total Incident Costs

As noted above, there are times when auxiliary information can be used to provide improved estimates of the total reported incident costs. Regression estimates are used to model the relationship between costs and the auxiliary variables, and the model is then applied to the entire population to obtain estimates for individual incidents, which are then summed to produce an estimated total cost. For this study, regression estimates were obtained separately for each stratum and the stratum totals were combined to produce an overall estimate.

Two difficulties in applying the regression method are when there are too few observations or very little variability in costs within a stratum. The former difficulty exists for water incidents, of which there were only two, and the latter difficulty exists for air incidents, where most of the incident costs were zero. As a result, regression estimates were not obtained for those two strata. Instead, the total reported incident costs for air and water incidents were determined using the stratified sampling estimates obtained above.

The HMIRS database contains a large number of variables that describe the incidents. Many of these incident characteristics may influence the overall cost for the incident and should be considered for inclusion in determining the regression estimate. Table 5 contains a list of variables from the Material and Container portions of the HMIRS database that were considered for inclusion in the regression estimate model. Descriptions of these variables are based on Form DOT 5800.1. For those variables in Table 5 that are categorical in nature, Table 6 shows the possible categories and defines new indicator variables that represent the various component options for each of the variables. Any classification variable with more than ten categories was excluded from the analysis.

Table 5. Potential Predictor Variables

Form DOT 5800.1 Question Number	HMIRS Variable Name	Description	Type of Variable
1	MODE	Mode of Transportation	Categorical
16	HAZSUB	Hazardous Substance Indicator	Indicator
18	RQUAN	Quantity Released	Continuous
18	RQMET	Release Occurred	Indicator
19	DEAD	Number of Deaths	Count
20	MJINJ	Number of Major Injuries	Count
21	MNINJ	Number of Minor Injuries	Count
22	NEVAC	Number Evacuated	Count
24	VAPOR	Vapor (Gas) Dispersion	Indicator
24	SEWER	Material Entered Waterway Sewer	Indicator
24	SPILL	Spillage	Indicator
24	FIRE	Fire	Indicator
24	EXPLO	Explosion	Indicator
24	ENVIR	Environmental Damage	Indicator
24	RNONE	No Consequences	Indicator
24	ROTH	Other Consequence	Indicator
25	CARGO	Cargo Tank	Indicator
25	VANTR	Van Truck / Trailer	Indicator
25	FLATT	Flat Bed Truck Trailer	Indicator
25	TCAR	Tank Car	Indicator
25	RCAR	Rail Car	Indicator
25	TOFC	TOFC/COFC	Indicator
25	PLANE	Aircraft	Indicator
25	BARGE	Barge	Indicator
25	SHIP	Ship	Indicator
25	VOTH	Other Vehicle	Indicator
26	PHASE	Transportation Phase	Categorical
27	LUSE	Land Use	Categorical
28	CTYPE	Community Type	Categorical
29	ACCDR	Accident/Derailment	Indicator
29A	SPEED	Estimated Speed	Continuous
29B	HYTPE	Highway Type	Component
29C	LANES	Number of Lanes	Count
--*	CAUSE	Cause of Incident	Categorical
--*	MISC1	Miscellaneous Code	Categorical
--*	MISC2	Miscellaneous Code	Categorical
32	NFAIL	Number of Failed Packages	Count
33	NSHIP	Number of Packages Shipped	Count
41a	VCOLL	Transport Vehicle Collision	Indicator
41b	VOVER	Transport Vehicle Overturn	Indicator
41c	OLOAD	Overloading/Overfilling	Indicator
41d	LOOSE	Loose Fittings/Valves	Indicator
41e	DEFCT	Defective Fittings/Valves	Indicator
41f	DROPD	Dropped	Indicator
41g	STRCK	Struck/Rammed	Indicator
41h	ILOAD	Improper Loading	Indicator
41i	BLOCK	Improper Blocking	Indicator
41j	CORRO	Corrosion	Indicator
41k	FATIG	Metal Fatigue	Indicator
41l	FRICT	Friction/Rubbing	Indicator
41m	HEAT	Fire/Heat	Indicator
41n	FREEZ	Freezing	Indicator
41o	VENT	Venting	Indicator
41p	VANDL	Vandalism	Indicator
41q	INCOM	Incompatible Materials	Indicator

Table 5. Potential Predictor Variables (Continued)

Form DOT 5800.1 Question Number	HMIRS Variable Name	Description	Type of Variable
41r	COTH	Other Contributing Factor	Indicator
42a	FRGHT	Other Freight	Indicator
42b	FLIFT	Forklift	Indicator
42c	NAIL	Nail/Protrusion	Indicator
42d	VEHCL	Other Transport Vehicle Cause	Indicator
42e	WATER	Water/Other Liquid	Indicator
42f	FLOOR	Ground/Floor/Roadway	Indicator
42g	OBSTC	Roadside Obstacle	Indicator
42h	NONE	No Object Caused Failure	Indicator
42i	OOTH	Other Object Caused Failure	Indicator
43a	PUNCT	Punctured	Indicator
43b	CRACK	Cracked	Indicator
43c	BURST	Burst/Internal Pressure	Indicator
43d	RIPPD	Ripped	Indicator
43e	CRUSH	Crushed	Indicator
43f	ABRAD	Rubbed/Abraded	Indicator
43g	RUPTD	Ruptured	Indicator
43h	HOTH	Other Failure	Indicator
44a	FORWD	End Forward Location	Indicator
44b	REAR	End Rear Location	Indicator
44c	RIGHT	Side Right Location	Indicator
44d	LEFT	Side Left Location	Indicator
44e	TOP	Top Location	Indicator
44f	BOTTM	Bottom Location	Indicator
44g	CENT	Center Location	Indicator
44h	AOTH	Other Location	Indicator
45a	MATRL	Basic Package Material	Indicator
45b	VALVE	Fitting Valve	Indicator
45c	CLOSE	Closure	Indicator
45d	CHIME	Chime	Indicator
45e	WELD	Weld/Seam	Indicator
45f	HOSE	Hose/Piping	Indicator
45g	INLIN	Inner Lining	Indicator
45h	WOTH	Other Package	Indicator

* DOT-Use-Only variable

There are more than 100 potential predictor variables shown in Tables 5 and 6, which is larger than the number of high-cost or rail incidents and approaches the number of highway incidents. As a result, not all of the predictors can be used. During some preliminary regression analyses, it was noted that "rare" predictors often had a major impact on the results, often to the detriment of the estimates (e.g., corrosion). As a result, to reduce the number of predictor variables and to remove variables that were overly sensitive, any predictor that occurred in fewer than five of the sampled incidents in a stratum was removed from the model for that stratum. Further analyses showed that there were still a large number of predictor variables that did not contribute to the prediction of the total costs in the regression analysis. In order to focus only on those variables that provided significant predictive ability, a stepwise regression analysis was performed for each stratum to select a set of variables that best predicted the total incident cost.

Table 6. Expansion of Categorical Variables to Indicator Variables

Categorical Variable	Codes	Interpretation	Indicator Variable
MODE	1	Air Mode	M_AIR
	4	Highway Mode	M_HWAY
	6	Rail Mode	M_RAIL
	7	Water Mode	M_WATER
CAUSE	10	Human Error Cause	C_HERR
	20	Package Failure Cause	C_PFAIL
	30	Accident/Derailment Cause	C_ACDR
	40	Other Cause	C_OTH
PHASE	261	En route Phase	P_ENRTE
	262	Loading Phase	P_LOAD
	263	Unloading Phase	P_UNLD
	264	Temp. Storage Phase	P_STOR
LUSE	271	Industrial Land	L_IND
	272	Commercial Land	L_COM
	273	Residential Land	L_RES
	274	Agricultural Land	L_AGR
	275	Undeveloped Land	L_UND
CTYPE	281	Urban Community	CM_URB
	282	Suburban Community	CM_SUB
	283	Rural Community	CM_RUR
HTYPE	291	Divided Highway	H_DIV
	292	Undivided Highway	H_UNDV
MISC1 MISC2	109	Evacuations	M_EVAC
	110	Container Failure (Release During Loading)	M_LOAD
	111	Human Error (Release During Loading)	M_LOAD
	112	Tank Failure (Release During Loading)	M_LOAD
	116	Train/Truck	M_TRTR
	118	Splash	M_SPLSH
	121	Rollover	M_ROLL
	122	Radioactive Release	M_RADIO
	128	Road Closure	M_RDCLO
	132	Loading Other	M_LOAD

Table 7 shows the variables (ordered by significance) that were selected by an initial stepwise regression procedure to be significant predictors of total incident cost within each of the three strata for which regression estimates were obtained. Interestingly, there were no variables that were included in more than one stratum. Also, no more than five variables were selected by the stepwise procedure for any of the strata.

Table 7. Significant Predictor Variables for Stratified Regression Estimate

High-cost	Highway	Rail
Material Entered Waterway/Sewer	Package Failure Cause	Other Location
Accident/Derailment	Human Error Cause	No Object Caused Failure
End Forward Location	Van Truck / Trailer	Spillage
Undivided Highway	Cracked	
	En route Phase	

Table 8 shows the results of the stratified regression estimation. The table contains regression estimates of the average incident cost by stratum and their standard errors, regression estimates of total stratum cost and their errors, confidence intervals for both average incident cost and total cost by stratum, and an estimate of the total incident costs over all strata. Included in the total estimate are the stratified sampling estimates for the air and water strata. Table 8 shows that the regression estimate of total reported incident costs is approximately $71 million, with a 95 percent confidence interval ranging from $62 million to $80 million. Appendix A shows the estimated regression equations relating total cost to the significant predictor variables.

There are several interesting results shown in Table 8 when compared to the stratified sampling results of Table 4. First, the estimated total costs in each of the three strata are higher using regression estimation than using stratified sampling estimates. In the case of rail incidents, the increase is notable. Secondly, the standard errors of the estimates are lower for regression estimates than for stratified sampling estimates. This can be explained partly by the fact that regression partitions variability into two components: a piece explained by the relationship between cost and its predictors, and a piece explained by random error. In the stratified sampling estimate, all variability is associated with random error. As a result, the standard error of estimates under stratified sampling should be larger than those under regression estimation.

Table 8. Stratified Regression Estimates for Average and Total Reported Incident Costs

Statistic	Stratum			
	High-cost	Highway	Rail	Total [1]
Number in Stratum	158	13,503	800	15,188
Number in Sample	59	129	62	260
Average Incident Cost ($)	266,325	1,975	3,130	NA
Std Error of Average Incident Cost ($)	26,866	115	807	NA
Lower 95% Confidence Bound for Average Incident Cost ($)	213,669	1,750	1,548	NA
Upper 95% Confidence Bound for Average Incident Cost ($)	318,982	2,220	4,712	NA
Stratum Total Cost ($)	41,546,765	26,666,234	2,504,046	70,938,089
Std Error of Stratum Total Cost ($)	4,191,058	1,552,043	645,826	4,518,389
Lower 95% Confidence Bound for Stratum Total Cost ($)	33,332,290	23,624,231	1,238,227	62,082,047
Upper 95% Confidence Bound for Stratum Total Cost ($)	49,761,239	29,708,237	3,769,865	79,794,132
HMIRS-Reported Stratum Total Cost ($)	32,788,896	15,999,911	825,193	49,821,192

1. Total includes Air and Water estimates from stratified sampling (see Table 4)

4.4 Stratified Estimate for Component Costs

The analyses discussed above sought to estimate the total reported incident costs over a one-year period either by expanding the information about total incident cost from a sample to the entire population of incidents or by modeling the total incident costs as a function of predictor variables. These estimates required that the total cost of a sampled incident be known. The total cost is the sum of five categories of costs, thus for the total cost to be known, each of the five component costs must be known. There are a number of sampled incidents where there was incomplete cost information, either in terms of component costs not being collected or in terms of collection of partial cost information for a component. Incidents with incomplete cost information can provide useful information for those categories where cost information is available.

Table 9 summarizes the available cost data for the individual categories within each stratum and overall. Of the individual categories, Material Loss has the fewest incidents with costs (310), which is an increase of 50 incidents over the 260 incidents used for the analyses based on only incidents for which data for all cost categories were available. The Property Damage component has the largest increase in available observations (94). The number of additional observations for modeling is significant enough to warrant obtaining separate cost estimates for each component and combining them to get an overall cost estimate for all categories combined.

Table 9. Number of Component Costs Collected, by Stratum and Overall

Stratum	Material Loss	Carrier Damage	Property Damage	Response Cost	Cleanup Cost	All Costs Present
High-cost	70	69	71	64	70	59
Highway	154	183	184	160	178	129
Rail	74	82	86	80	81	62
Air	8	9	9	9	9	8
Water	4	4	4	2	4	2
Total	310	347	354	315	342	260

Summing the estimated costs for individual categories adds a level of complexity to the estimates beyond the requirement of five separate estimates. Incidents in different strata are independent, thus the variance of the sum of the costs across strata is simply the sum of the stratum variances. Component costs, on the other hand, likely are not independent. In fact, there may be significant correlations between the costs for different categories. For example, if material loss costs are high, cleanup costs are more likely to be high, indicating a positive correlation between those two categories. The result of summing correlated costs is that the variance of the sum must be adjusted for the correlation between cost components. Equations (6) and (7) above show the modification to the variance formula for correlated costs.

Table 10 shows the correlation coefficients calculated between each pair of categories. Correlation coefficients are shown for each of the three strata with sufficient data (high-cost, highway, rail) as well as combined across all strata. For each correlation coefficient, the number of observations used in its calculation is shown in parentheses.

Examination of Table 10 indicates that the correlation coefficient between cost categories varies among the strata. Correlations between categories for rail incidents are generally quite high in comparison to the other two strata. There appear to be no general trends in correlation coefficients other than the fact that nearly all correlation coefficients are positive. The one exception – between property damage and carrier damage in highway incidents – is effectively equal to zero.

Table 10. Correlation Coefficients (Number of Incidents) Between Cost Categories (No Regression Model), by Stratum and Overall

Component	Material Loss	Carrier Damage	Property Damage	Response Cost
High-cost				
Carrier Damage	0.003 (67)			
Property Damage	0.078 (68)	0.576 (68)		
Response Cost	0.012 (61)	0.226 (61)	0.244 (64)	
Cleanup Cost	0.043 (68)	0.369 (68)	0.360 (69)	0.471 (63)
Highway				
Carrier Damage	0.066 (153)			
Property Damage	0.200 (154)	-0.013 (183)		
Response Cost	0.416 (132)	0.378 (159)	0.013 (160)	
Cleanup Cost	0.319 (151)	0.146 (177)	-0.027 (178)	0.310 (156)
Rail				
Carrier Damage	0.759 (70)			
Property Damage	0.761 (74)	0.964 (82)		
Response Cost	0.013 (70)	0.172 (76)	0.010 (80)	
Cleanup Cost	0.817 (70)	0.954 (77)	0.976 (81)	0.017 (75)
All Strata Combined				
Carrier Damage	0.167 (302)			
Property Damage	0.128 (308)	0.569 (346)		
Response Cost	0.163 (273)	0.471 (307)	0.308 (315)	
Cleanup Cost	0.145 (301)	0.496 (335)	0.402 (341)	0.581 (305)

Table 11 contains the stratified sampling estimates for cost components, including estimates for total costs obtained by summing the component costs. This table includes estimates for the mean and total component cost for each stratum, as well the total component cost across all strata and the total stratum cost across all categories. For the latter case, no average incident cost is presented. Table 11 shows that stratified sampling estimate of total reported incident costs, based on component cost estimates, is approximately $118 million. While not shown in Table 11, the 95 percent confidence interval for the costs ranges from $84 million to $152 million.

Table 11. Stratified Sampling Estimates for Component Costs

Stratum	Number in Stratum	Number in Sample	Average Incident Cost ($)	Std Error of Average Incident Cost ($)	Stratum Total Cost ($)	Std Error of Stratum Total Cost ($)
Material Loss						
High-cost	158	70	10,857	30,860	1,693,682	427,226
Highway	13,503	154	134	311	1,804,287	336,098
Rail	800	74	21	72	17,088	6,348
Air	718	8	0	0	0	0
Water	11	4	10	20	110	88
Total	15,188	310	NA	NA	3,515,466	543,621
Carrier Damage						
High-cost	158	69	69,957	89,169	10,913,266	1,250,577
Highway	13,503	183	1,277	10,811	17,243,124	10,717,959
Rail	800	82	232	1,310	185,889	109,620
Air	718	9	0	0	0	0
Water	11	4	150	300	1,650	1,316
Total	15,188	347	NA	NA	28,343,929	10,791,228
Property Damage						
High-cost	158	71	13,139	59,980	2,049,714	819,683
Highway	13,503	184	3	23	34,858	22,382
Rail	800	86	268	2,480	214,419	202,122
Air	718	9	0	0	0	0
Water	11	4	0	0	0	0
Total	15,188	354	NA	NA	2,298,991	844,532
Response Cost						
High-cost	158	64	263,150	323,090	41,051,383	5,174,223
Highway	13,503	160	1,870	8,204	25,245,529	9,706,339
Rail	800	80	2,304	7,721	1,843,452	753,401
Air	718	9	69	128	49,632	32,290
Water	11	2	15,583	21,973	171,413	154,591
Total	15,188	315	NA	NA	68,361,409	11,026,248
Clean-up Cost						
High-cost	158	70	72,496	154,043	11,309,365	2,132,569
Highway	13,503	178	292	1,073	3,943,883	1,079,278
Rail	800	81	387	2,592	309,798	218,447
Air	718	9	0	0	0	0
Water	11	4	0	0	0	0
Total	15,188	342	NA	NA	15,563,045	2,400,086
Total						
High-cost	NA	NA	NA	NA	67,017,377	6,612,993
Highway	NA	NA	NA	NA	48,271,680	16,010,547
Rail	NA	NA	NA	NA	2,570,645	881,342
Air	NA	NA	NA	NA	49,632	32,290
Water	NA	NA	NA	NA	173,173	154,597 [1]
Total	NA	NA	NA	NA	118,082,507	17,345,634

1. The standard error for water incidents was estimated using the sum of the variances for individual components, without considering correlations, which were not estimable.

4.5 Stratified Regression Estimate for Cost Components

The extension of the stratified regression approach to component costs is quite natural. In fact, different variables would be expected to influence costs within each component, and the significant variables may also differ between the different strata. Thus, the stratified regression approach was used to estimate component costs within each stratum. As was the case based on total cost, the air and water strata were excluded from this analysis. To estimate total reported incident costs for these two strata, the stratified sampling results for component costs, as shown in Table 11, were used.

Table 12 shows the variables (ordered by significance) that were significant predictors of the component costs for each of the three strata (high-cost, highway, rail). In general, there are more significant predictor variables in each of the stratum-component models than there were in the models for total incident costs within stratum. There are also several predictors that appear more consistently within categories across strata. For example, an indicator variable noting that the incident was caused by human error appears in the carrier damage component for both highway and rail. Similarly, there are several predictors that appear in multiple categories within a stratum. For example, a release occurring in a forward location within a package appears in the material loss, carrier damage, and property damage categories for high-cost incidents.

The fact that there are predictors that appear in more than one component within a stratum is indicative of correlations between the component costs. Table 13 shows the partial correlations between component costs that were estimated within each of the three strata. Partial correlations were obtained after multivariate models were fitted to the component cost data including all significant predictor variables and represent the correlation between the component costs after adjusting for the predictors. Table 13 shows that there are several more negative correlations than there were prior to adjustment for the predictors. This table also shows that the partial correlations vary among the three strata, as did the correlations shown in Table 10.

Table 12. Significant Predictors for Cost Components (by Stratum)

Material Loss	Carrier Damage	Property Damage	Response Cost	Cleanup Cost
High-cost				
Fire/Heat	Basic Package Material	Evacuations	Rail Mode	Material Entered Waterway/Sewer
End Forward Location	End Forward Location	End Forward Location	Material Entered Waterway/Sewer	No Object Caused Failure
Bottom Location	Accident/Derailment	Transport Vehicle Collision	Ground/Floor	Suburban Community
Cracked	Road Closure	Number Evacuated	Transport Vehicle Overturn	
Divided Highway	Unloading Phase	End Rear Location	Fire	
Other Cause	Undivided Highway	Estimated Speed	Commercial Land	
	Cracked		Side Left Location	
Highway				
Punctured	Human Error Cause	Release During Loading	Cracked	Van Truck / Trailer
Hazardous Substance	Package Failure Cause	Urban Community	Human Error Cause	Closure
En route Phase	Spillage	Other Failure	Package Failure Cause	Commercial Land
Struck/Rammed	Rural Community	Other Location	Improper Blocking	
		Van Truck / Trailer	Dropped	
			Other Failure	
			Loose Fittings/Valves	
			Burst	
			Other Failure	
			Improper Loading	
Rail				
Other Location	Human Error Cause	Other Location	Tank Car	Other Location
Other Cause	Package Failure Cause	Basic Package Material	Suburban Community	No Object Caused Failure
No Object Caused Failure	Hazardous Substance	No Object Caused Failure	Top Location	
	Top Location	Other Failure	En route Phase	

Table 13. Partial Correlations Between Component Costs from Regression Model (by Stratum)

	Material Loss	Carrier Damage	Property Damage	Response Cost
High-cost				
Carrier Damage	-0.2752			
Property Damage	-0.1568	0.5702		
Response Cost	0.3006	0.1548	0.0623	
Cleanup Cost	-0.0089	0.4595	0.4840	0.379
Highway				
Carrier Damage	-0.3533			
Property Damage	0.30630	-0.0034		
Response Cost	0.2379	-0.7311	0.0882	
Cleanup Cost	0.1576	-0.4337	-0.0893	0.3083
Rail				
Carrier Damage	0.2114			
Property Damage	-0.0361	-0.0204		
Response Cost	0.0967	0.6971	0.0448	
Cleanup Cost	-0.0387	0.1724	0.0422	0.0683

Table 14 contains the stratified regression estimates for cost components, including estimates for each combination of stratum and component, total component estimates, and totals within each stratum. Average costs are included for each stratum-component combination, but not for totals across categories, across strata, or overall. Table 14 shows that stratified sampling estimate of total reported incident costs, based on component cost estimates, is approximately $78 million. While not shown in Table 14, the 95 percent confidence interval for the costs ranges from $67 million to $89 million. This estimate is significantly higher than the stratified sampling and stratified regression estimates based on total incident costs. This might be explained by the presence of higher component costs for those incidents that do not have all categories completed than for those incidents where all five component costs are present. Note that there are large discrepancies in high-cost, highway, and rail incidents compared to the estimate obtained in the stratified sampling and shown in Table 4. Appendix A shows the estimated regression equations relating total cost to the significant predictor variables. Appendix B presents the results of an analysis comparing serious incidents, as defined by PHMSA, in the sampled reported incidents to those in the high-cost stratum.

Table 14. Stratified Regression Estimates for Component Cost

Stratum	Number in Stratum	Number in Sample	Average Incident Cost ($)	Std Error of Average Incident Cost ($)	Stratum Total Cost ($)	Std Error of Stratum Total Cost ($)
Material Loss						
High-cost	158	70	13,221	1,871	2,062,467	291,889
Highway	13,503	154	161	25	2,177,715	342,865
Rail	800	74	23	8	18,016	6,109
Air [1]	718	8	0	0	0	0
Water [1]	11	4	10	20	110	88
Total	15,188	310	NA	NA	4,258,308	450,326
Carrier Damage						
High-cost	158	69	61,876	6,045	9,652,616	942,982
Highway	13,503	183	1,414	389	19,096,693	5,243,108
Rail	800	82	332	36	265,533	28,441
Air	718	9	0	0	0	0
Water	11	4	150	300	1,650	1,316
Total	15,188	347	NA	NA	29,016,493	5,327,308
Property Damage						
High-cost	158	71	13,621	4,284	2,124,841	668,297
Highway	13,503	184	3	2	39,236	21,199
Rail	800	86	450	236	359,600	188,821
Air	718	9	0	0	0	0
Water	11	4	0	0	0	0
Total	15,188	354	NA	NA	2,523,676	694,783
Response Cost						
High-cost	158	64	83,568	7,404	13,036,617	1,155,030
Highway	13,503	160	850	201	11,472,820	2,717,718
Rail	800	80	1,220	177	976,207	141,409
Air	718	9	69	128	49,632	32,290
Water	11	2	15,583	21,973	171,413	154,591
Total	15,188	315	NA	NA	25,706,688	2,960,578
Clean-up Cost						
High-cost	158	70	78,260	10,761	12,208,524	1,678,709
Highway	13,503	178	267	75	3,601,335	1,012,910
Rail	800	81	465	250	371,921	200,189
Air	718	9	0	0	0	0
Water	11	4	0	0	0	0
Total	15,188	342	NA	NA	16,181,781	1,970,818
Total						
High-cost	NA	NA	NA	NA	39,085,065	2,854,029
Highway	NA	NA	NA	NA	36,387,798	4,877,853
Rail	NA	NA	NA	NA	1,991,278	324,043
Air	NA	NA	NA	NA	49,632	32,290
Water	NA	NA	NA	NA	173,173	154,597
Total	NA	NA	NA	NA	77,686,946	5,662,939

1. Air and Water results are from stratified sampling estimates, not regression estimates

5.0 DISCUSSION

Four different statistical methods were used to estimate the total financial consequence of reported hazardous material incidents over a one-year period. While all four methods are valid, there are differences among them that lead to discrepancies in their results. Table 15 shows the estimated total costs for each of the four methods. Note that in the HMIRS database, for the 15,188 records that comprise the year-long period to which the estimates apply, the total reported incident cost was $49,821,192. Of the four statistical estimates of the total reported incident costs, the stratified sampling estimate based on total incident costs was the smallest, while the stratified sampling estimate based on component costs was the highest. Regression estimation significantly reduced the variability of the estimates, both for total-cost estimation and component-cost estimation. Three of the four estimates are in fairly close agreement, with only the stratified estimate based on component costs significantly deviating from the other estimates.

Table 15. Comparison of Total Reported Incident Cost Estimates (in Dollars)

Method	Estimated Total Cost	Std. Error	Lower 95% Confidence Bound	Upper 95% Confidence Bound
Stratified, total cost	68,361,409	11,026,248	46,749,963	89,972,855
Regression, total cost	70,938,089	4,518,389	62,082,047	79,794,132
Stratified, component costs	118,082,507	17,345,634	84,085,065	152,079,949
Regression, component costs	77,686,946	5,662,939	66,587,586	88,786,306

It is reasonable, before drawing conclusions about the estimates, to examine their validity. This requires an assessment of the quality of the data and the statistical methods. There are several issues regarding the data that may affect their quality. These include:

- failure to meet sample size requirements;
- presence and treatment of missing cost data;
- discrepancies between HMIRS data and sample data;
- treatment of air and water incidents;
- validity of the predictor variables;
- potential missing predictors;
- regression model validity;
- unusual values of predicted total incident costs; and
- alternative models.

These issues are discussed in the paragraphs that follow.

Failure to Meet Sample Size Requirements

Table 2 shows the number of reported hazardous material incidents during the target year that formed the basis of the sample selection. Table 2 also shows the number of incidents sampled in each of the five strata, as well as the number of incidents for which the revised cost data were obtained. The original sampling plan called for a sample of 400 incidents, with an additional 100 incidents being added to account for non-response. Table 2 indicates that the targeted number of incidents was not met in any of the five strata. Overall, only 65 percent of the targeted number of samples was completed, although there were a number of sampled incidents for which partial data were collected.

The failure to meet the targeted sample size does not invalidate the results, however. The result of having too few completed incidents is that the targeted precision of the estimates was not met. For the study, the actual precision of the estimates was calculated, and confidence intervals were calculated based on the precision that was attained.

As a result of failing to meet the target, modifications to any data collection plans for the future should include an increase in the expected non-response rate. In the future, the number of incidents sampled should be increased so that the target of 400 incidents is met, or the target number of incidents should be decreased to match the results found in the current study.

Presence and Treatment of Missing Cost Data

During interviews, the organizations responsible for providing incident cost information to HMIRS were asked to review their records and provide revised cost information. The costs were broken down into several categories, and the organizations were asked to provide cost information by component. In a few cases, organizations were able to provide revised costs for some of the categories but did not provide costs for other categories. Sometimes, other sources were identified for these missing costs. These included shippers, cleanup companies, emergency responders, and others. In some cases, these additional sources were not able to provide complete information. As a result, the total incident cost – which is equal to the sum of the costs per component – could not be determined, and such incidents could not be used for the statistical modeling of total incident costs. However, partial cost information obtained from the reporting organizations provided useful data about component costs and should be used in some capacity. To take full advantage of the partial data that were available, models were fitted not only to the total incident costs, but also to each of the component costs. Those incidents with only partial cost information were excluded from the analysis for total incident cost, but they were included for those categories for which they had revised data.

As noted in Table 15, there was a significant discrepancy in the estimate of total reported incident costs for the stratified sampling method when component costs were used rather than total incident cost. This discrepancy is probably due to differences in the incidents for which partial information was obtained. The incidents with partial costs may have had higher component costs on average than the incidents for which all component costs were obtained. This would have resulted in the significantly larger estimate obtained when component costs were used. Further investigation should determine whether the inclusion or exclusion of incidents with partial data leads to a bias in the results.

Discrepancies between HMIRS Data and Sample Data

The HMIRS database divides total incident costs into five categories:

- Product lost (PLDAM);
- Carrier damage (CADAM);
- Public/Private property damage (PPDAM);
- Decontamination costs (DCDAM); and
- Other costs (OTDAM).

In interviewing the incident reporting organizations, revised cost estimates were requested for five categories that did not directly correspond to the five HMIRS categories. Revised costs are available for

- Material loss;
- Carrier damage;
- Property damage;
- Response costs; and
- Clean-up costs.

While four of the categories appear to be equivalent, the “Other costs” and “Response Costs” categories do not obviously correspond.

Any potential discrepancy between the categories did not cause any analysis problems. The analysis was based, first of all, on total incident costs, which was defined as the sum of the costs for the five categories of revised costs. Secondly, the component-cost analyses focused on the categories for revised costs used in the study. Finally, comparisons of total-incident and component-sum results did not include any of the cost information from the HMIRS database. On the other hand, a detailed comparison of the discrepancies between HMIRS and study component costs could not be done because of the differences in component definitions. In addition, if future modeling efforts used HMIRS component costs as potential predictors, there could be problems due to component definitions. This problem should be eliminated, however, for data reported after January 1, 2005 as the cost categories on the revised Form DOT 5800.1 will be aligned with those used in this study.

Treatment of Air and Water Incidents

In the presentation of the results of the statistical analyses, it was noted that regression estimates for air and water incidents could not be obtained. As a result, the stratified regression estimates of total reported incident costs included air and water estimates obtained using stratified sampling estimates. Combining estimates using two different methods does not cause a problem with the results, because stratum estimates are independent and their variances can be added.

Nonetheless, it should be noted that the estimation method does differ for different pieces of the total.

Validity of Predictor Variables

An examination of the HMIRS database shows that nearly every record contains data for all of the predictor variables that were used in the regression analyses. However, it is not known to what extent the predictor data in the HMIRS are correct. During the study, only cost information was requested, so any potential incorrect or missing data for the predictor variables were not corrected. The presence of incorrect predictors can potentially be great, especially when they are associated with rare events (e.g., deaths). While it was not considered prior to conducting the study, it may have been useful to verify the predictor variables that were reported in the HMIRS database. Alternatively, the use of third-party data collection and verification also may have been useful in validating the predictor variables.

Potential Missing Predictors

The set of predictor variables that were used in the regression analysis included many of the variables contained in the Material and Container tables of the HMIRS database. There were several variables, however, that were not included that might prove to be significant predictors of incident costs. The HMIRS database notes the location where the incident occurs, but this information was not used in the regression analysis. Thus, if incident costs differ by geographical region, that relationship was not included in the model. Another set of potential predictors that was not incorporated into the model relates to the chemical properties of the hazardous material. It would seem likely that the incident costs may be strongly related to the chemical characteristics, so future analyses may want to include variables that capture the chemical properties. The variable CMCL in the HMIRS database might prove to be a good initial proxy for this information. It was not used in the current analyses because there were more than 10 categories. There may be other predictors that also were not included. A review of the data available in the HMIRS database should be undertaken to ensure that all potentially important predictors can be included in future models.

Regression Model Validity

To perform a valid regression analysis, several assumptions must be met. First, the regression model includes an error term that is assumed to have a normal distribution with mean zero and constant variance. Error terms are also assumed to be independent between incidents. These two assumptions were examined using post-regression diagnostic plots, including plots of residuals (which are the differences between observed and predicted costs and which represent estimates of the errors) versus predicted and observed incident costs. These plots showed that the variance could be assumed to be constant rather than being a function of the cost. The diagnostic plots also showed that incident “error” appeared to be independent. These diagnostic plots, along with plots comparing residuals and predicted values to the predictor variables, also were used to assess model validity. The plots were examined for the presence of patterns that might indicate non-linear relationships between costs and predictors. In several cases, there were pronounced patterns in residuals, but these were generally the result of the small number of variables entered into the model rather than non-linear relationships.

One additional assumption of regression analysis is that any predictions made about incident costs should be restricted to the bounds of the data used to quantify the relationships between the costs and predictors. This general assumption applies to the values of the predictor variables. However, in this case, care should be exercised in applying the results outside of the temporal bounds of the data. That is, the estimated reported incident costs are applicable to incidents occurring between November 1, 2002, and October 31, 2003. Should estimates be required for other time intervals, caution should be used in interpreting the results. Using a greater time period for modeling and using a model where the predictor variables are chosen using a non-stepwise procedure should produce a more durable model that is more applicable beyond the temporal limits of the data used to produce it.

Unusual Values of Predicted Total Incident Costs

In examining the regression diagnostic plots, it was noted that some of the predicted costs were less than zero. This is something that obviously cannot be true for individual incidents. However, the objective of the regression estimate was to determine the general relationship between costs and predictor variables and to apply those relationships to all of the incidents over a year’s length. Thus, with the level of specificity currently available with the limited sample size, individual incident cost estimates are not meaningful, although the overall incident cost estimates are meaningful because they are the sum over all incidents. As additional incidents are sampled in future efforts, the estimates for an individual incident’s cost will become more meaningful.

Alternative Models

There are alternative regression models that might be considered as replacements for the linear regression models used in this analysis. Some of these models might involve transforming the costs to avoid possible negative values. Other of these alternative methods would add “prior” information that might improve the estimates. Two such models were considered for this analysis.

To eliminate the possibility of a negative cost and to better model the “skewed” distribution of costs, logarithms of the costs could be taken and regression models be fitted to them. To avoid problems associated with incidents with no costs, $1 would be added to each incident before taking the logarithms and subtracted from predicted incident costs after the model was fitted. The predicted values, after transforming back to the original scale, would be summed to provide an estimate of the total incident costs. Determining the precision of these estimates requires first-order Taylor-series expansion that incorporates the errors of the predicted log-costs and the exponentiated mean cost.

There are problems associated with this method of estimation as well as benefits. The obvious benefit is that negative costs are not allowed (or limited to less than $1). Several problems with this method include:

- variances of the total costs are more difficult to obtain;
- variances of the total costs are usually significantly greater than their counterparts when no logarithmic transformation is taken; and
- low-cost incidents are usually over-estimated and high-cost incidents are usually underestimated, although there are cases where extreme observations are made even more extreme when using logarithms.

On the whole, the cost estimates obtained using logarithmic transformations are less well-behaved and produce greater cost estimates, with higher variability, than when no transformations are taken.

A second alternative regression method, which also addresses the problem of negative cost estimates, is “tobit” regression. This econometric analysis method was developed specifically to address data where the response variable is censored. Tobit regression analysis decomposes parameter estimates into two parts, one to address the positive probability that a response is equal to zero, and the other to address means conditional upon them being greater than zero. Use of this model would result in forcing all estimated incident costs to be greater than or equal to zero. This model was not employed because it is designed for use with data that are censored rather than data that have a positive probability of some specific value. In actuality, incident costs cannot be negative, so they are not, in fact, censored. As a result, the tobit model has not been considered to be appropriate for incident costs.

One variable that was not considered in the cost models was the estimated incident cost that was available in the HMIRS database. It stands to reason that even if it is not very accurate, the preliminary cost estimate will provide some measure of what the total final cost is. By using it as a predictor, it is possible that the model might be improved. Because of the discrepancy in cost categories between HMIRS and the study data, use of HMIRS reported costs should be limited to estimates involving total cost only.

The stratified regression approach using total cost was applied, adding RPDAM (reported total incident cost) as a potential predictor. For all three strata (high-cost, highway, rail), it was a significant predictor of total incident cost. In the case of highway incidents, it improved the regression model dramatically (as measured by the model R^2 value). However, in the other two cases, the model was not greatly improved. The estimated total reported incident costs per stratum did not change greatly (i.e., increased by approximately $4 million). In addition, the precision of the estimates decreased slightly (approximately 10 percent) for each of the strata. This indicates that adding the reported costs does not improve the estimates greatly.

6.0 CONCLUSIONS AND RECOMMENDATIONS

Four statistical estimates of total reported incident costs were obtained. Three of the estimates were in fairly good agreement, while the fourth was significantly larger. For the three similar estimates, the total reported incident costs estimated for the period of November 1, 2002, through October 31, 2003, were approximately $70 million to $75 million, which was about 40 to 45 percent higher than the total reported incident costs in the HMIRS database.

Based on the strengths and weaknesses of each of the estimation methods and the quality of the results achieved, it is recommended that the results of the regression estimation based on cost components be used to estimate the total reported incident costs. This estimate is recommended because it makes the maximum use of the limited data that were available and it produces the estimate with the smallest variability. The stratified estimate based on total cost, while providing a reasonable estimate, did not make use of any relationships between the costs and the predictor variables, nor did it use all the information available. The stratified regression estimate also was reasonably consistent with the other estimates, but it also did not make full use of the data available. The stratified sampling estimate based on component costs used all of the cost information available, but it did not use any predictor variables to reduce the variability. In addition, a few discrepancies between the sample and the population resulted in a much larger estimate than in the other three methods.

Based on the results found and reported here, and also based on information that was unavailable or that could not be obtained for this analysis, the following recommendations are made for future studies.

- Future data collection to improve cost estimates should include verification of the predictor variables in addition to collection of revised costs. While there is no evidence that the predictor variables are not correct, it would seem likely that some mistakes have been made in completing and recording the data.

- An additional study should be undertaken to examine the extent to which the models found in this analysis are applicable beyond the temporal extent of this study. This additional study should be similar to the current study, with statistical analyses examining not only what is the best model for the cost data for another time period, but the extent to which the model for the time period covered in this study can predict costs in the new time period.

- Should additional studies be undertaken, or should data collection be continued as a means to obtain “revised” incident costs, it would be useful to include some of the explanatory variables that could not be included in this study. Chief among these are chemical characteristics data for the hazardous materials involved in the incident. Other variables that could be included might be geographic region and packaging material used. A review of all data available in the HMIRS database should be undertaken to ensure that all potentially important predictors can be included in future models.

- As there was a significant discrepancy in the estimate of total incident costs for the stratified sampling method when component costs were used rather than total incident costs, further investigation should determine whether the inclusion or exclusion of incidents with partial data leads to a bias in the results.

- The same sampling plan should be used for future data collection. However, 49 CFR 171.21 should be cited to improve the response rate. This regulation requires that any entity that reports an incident to the HMIRS must provide timely assistance to the DOT for any study involving that incident. The resulting increase in incidents for which data are obtained will improve the predictive ability of the statistical models.

- A detailed comparison of the discrepancies between reported HMIRS component costs and those identified in the study can be completed after sufficient data are reported to PHMSA via new incident reporting form that goes into effect on January 1, 2005. The new form contains the same cost categories as those used in the study.

APPENDIX A

STATISTICAL REGRESSION EQUATIONS FOR PREDICTING INCIDENT COSTS

This appendix contains detailed results of the regression model fitting for estimating total reported incident costs. Table A-1 and A-2 contain the estimated regression parameters that produce predicted incident costs for the various strata. Table A-1 shows the regression estimates when the total incident cost is predicted, and Table A-2 shows the regression estimates when individual component costs are predicted. Tables A-1 and A-2 also provide information about the fit of the regression models. The "fit" information includes the model R^2, which measures the proportion of the variability in the data that is explained by the predictor variables, and the square root of the mean square error, which represents the overall unknown variability in the model.

Tables A-1 and A-2 are useful in predicting costs for any incident given the values of the predictors. The cost is predicted by starting with the intercept term, and adding or subtracting the remaining amounts in the "Estimate" column if the "Term" variable is for a characteristic of the incident. Thus, for a highway incident caused by human error, the cost would be equal to $64,972 - $63,200 = $1,772. There are two exceptions to this method for determining the predicted cost. For the predictor variables "Speed" and "No. Evacuated," the estimate is multiplied by the speed or the number of persons evacuated, and that product is added to the prediction.

Figures A-1 through A-3 show the results of the prediction model in terms of plots of the predicted costs versus the actual costs. Figure A-1 show these plots for high-cost incidents, Figure A-2 shows them for highway incidents, and Figure A-3 shows them for rail incidents. In all three cases, the figure is divided into six plots. The upper left plot shows the predicted versus actual total costs, while the remaining five plots show the predicted versus actual costs for the five cost categories.

Table A-1. Regression Equations for Predicting Total Incident Costs

High-cost		Highway		Rail	
Term	**Estimate (p-value)**	**Term**	**Estimate (p-value)**	**Term**	**Estimate (p-value)**
R^2	0.3892	R^2	0.9754	R^2	0.3013
Root MSE	261,699	Root MSE	1,312	Root MSE	6,618
Intercept	60,824	Intercept	64,972	Intercept	-4,250
Material Entered Waterway/Sewer	384,349 (0.0006)	Package Failure Cause	-63,369 (<0.0001)	Other Location	15,912 (<0.0001)
Accident/Derailment	260,279 (0.0017)	Human Error Cause	-63,200 (<0.0001)	No Object Caused Failure	4,597 (0.0260)
End Forward Location	231,452 (0.0208)	Van Truck / Trailer	-1,187 (0.0007)	Spillage	2,789 (0.1325)
Undivided Highway	-209,760 (0.0230)	Cracked	1,016 (0.0499)		
		En route Phase	1,063 (0.0600)		

Table A-2. Regression Equations for Predicting Component Costs

High-cost				Highway				Rail			
Term	Estimate	t	p-value	Term	Estimate	t	p-value	Term	Estimate	t	p-value
Material Loss											
R^2	0.5738			R^2	0.1526			R^2	0.2680		
Root MSE	21,084			Root MSE	290			Root MSE	63		
Intercept	-4,341	-1.26	0.2114	Intercept	82	3.04	0.0028	Intercept	-24	-1.47	0.1474
Fire/Heat	61,829	4.60	<0.0001	Punctured	328	3.43	0.0008	Other Location	88	3.39	0.0012
End Forward Location	21,440	2.91	0.0050	Hazardous Substance	287	3.02	0.0030	Other Cause	41	2.06	0.0430
Bottom Location	24,965	2.87	0.0055	En route Phase	196	2.02	0.0447	No Object Caused Failure	36	2.00	0.0494
Cracked	28,504	2.81	0.0066	Struck/Rammed	-181	-1.85	0.0670				
Divided Highway	16,039	2.71	0.0087								
Other Cause	-35,874	-2.53	0.0141								
Carrier Damage											
R^2	0.4900			R^2	0.7660			R^2	0.9360		
Root MSE	67,237			Root MSE	5,289			Root MSE	340		
Intercept	83,670	4.14	0.0001	Intercept	67,571	17.31	<0.0001	Intercept	11,522	33.90	<0.0001
Basic Package Material	-87,686	-4.66	< 0.0001	Human Error Cause	-75,058	-22.80	<0.0001	Human Error Cause	-11,333	-32.36	<0.0001
End Forward Location	99,139	3.89	0.0003	Package Failure Cause	-74,541	-21.61	<0.0001	Package Failure Cause	-11,162	-31.59	<0.0001
Accident/Derailment	61,938	2.95	0.0045	Spillage	7,273	2.20	0.0292	Hazardous Substance	-11,522	-23.97	<0.0001
Road Closure	-63,637	-2.81	0.0067	Rural Community	2,208	1.95	0.0533	Top Location	-191	-2.10	0.0392
Unloading Phase	-77,442	-2.48	0.0159								
Undivided Highway	-52,687	-2.35	0.0221								
Cracked	-72,505	-2.23	0.0292								

Table A-2. Regression Equations for Predicting Component Costs (Continued)

High-cost				Highway				Rail			
Term	Estimate	t	p-value	Term	Estimate	t	p-value	Term	Estimate	t	p-value
Property Damage											
R^2	0.3922			R^2	0.1615			R^2	0.1684		
Root MSE	48,902			Root MSE	21			Root MSE	2,317		
Intercept	7,614	0.90	0.3723	Intercept	-13	-1.83	0.0683	Intercept	730	0.61	0.5410
Evacuations	71,289	3.80	0.0003	Release During Loading	28	4.11	<0.0001	Other Location	3,840	3.99	0.0001
End Forward Location	53,692	2.77	0.0074	Urban Community	7	2.14	0.0335	Basic Package Material	-2,025	-2.14	0.0354
Transport Vehicle Collision	47,021	2.58	0.0122	Other Failure	13	2.47	0.0146	No Object Caused Failure	1,241	1.92	0.0578
Number Evacuated	-85	-2.55	0.0131	Other Location	-9	-1.84	0.0676	Other Failure	-1,713	-1.46	0.1480
End Rear Location	-39,323	-2.07	0.0422	Van Truck / Trailer	11	1.64	0.1020				
Estimated Speed	-540	-2.06	0.0437								
Response Cost											
R^2	0.5186			R^2	0.4606			R^2	0.1985		
Root MSE	77,131			Root MSE	2,561			Root MSE	1,667		
Intercept	-25,489	-1.28	0.2051	Intercept	12,970	6.74	<0.0001	Intercept	1,952	2.14	0.0356
Rail Mode	155,945	5.58	<0.0001	Cracked	4,832	5.45	<0.0001	Tank Car	-1,214	-2.24	0.0278
Material Entered Waterway/Sewer	119,009	3.77	0.0004	Human Error Cause	-12,533	-6.63	<0.0001	Suburban Community	-816	-2.12	0.0373
Ground/Floor	86,006	3.46	0.0010	Package Failure Cause	-13,157	-6.87	<0.0001	Top Location	-946	-2.03	0.0457
Transport Vehicle Overturn	55,561	2.59	0.0122	Improper Blocking	4,259	4.55	<0.0001	En route Phase	1,323	1.65	0.1041
Fire	59,976	2.18	0.0355	Dropped	-3,800	-4.34	<0.0001				
Commercial Land	48,091	2.14	0.0369	Other Failure	-2,190	-3.65	0.0004				
Side Left Location	-46,873	-1.68	0.0985	Loose Fittings/Valves	-2,663	-3.59	0.0004				
				Burst	3,542	3.53	0.0005				
				Other Failure	2,537	3.34	0.0011				
				Improper Loading	-1,303	-2.01	0.0468				

Table A-2. Regression Equations for Predicting Component Costs (Continued)

High-cost				Highway				Rail			
Term	Estimate	t	p-value	Term	Estimate	t	p-value	Term	Estimate	t	p-value
Cleanup Cost											
R^2	0.4073			R^2	0.1341			R^2	0.1812		
Root MSE	121,259			Root MSE	1,007.47			Root MSE	2,376		
Intercept	39,255	2.07	0.0420	Intercept	823	3.76	0.0002	Intercept	-1,234	-2.00	0.0487
Material Entered Waterway/Sewer	262,928	5.36	<0.0001	Van Truck / Trailer	-1,087	-4.56	<0.0001	Other Location	4,265	4.09	0.0001
No Object Caused Failure	-102,372	-2.68	0.0092	Improper Closure	571	3.50	0.0006	No Object Caused Failure	1,664	2.45	0.0166
Suburban Community	76,782	2.39	0.0198	Commercial Land	338	2.17	0.0311)				

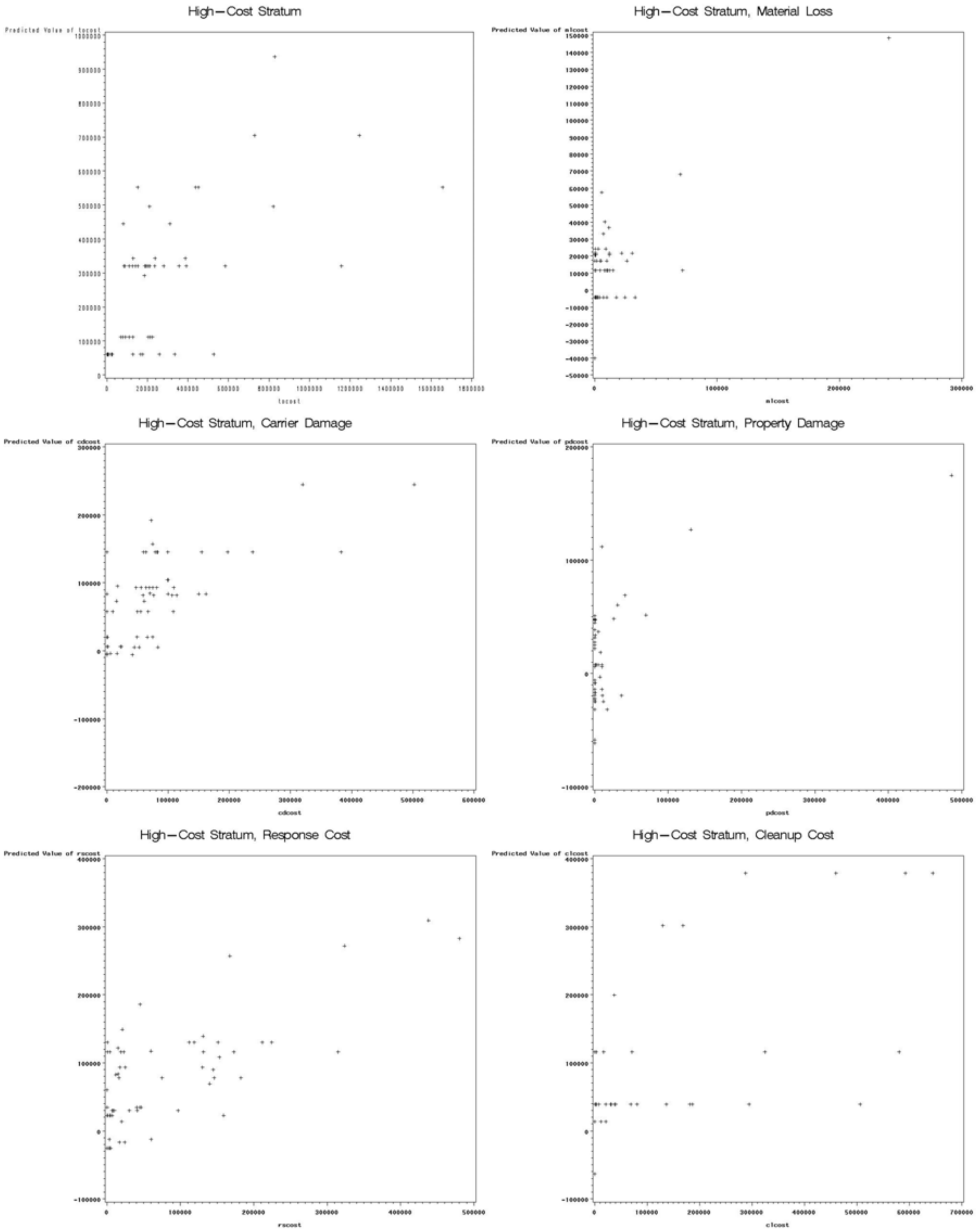


Figure A-1. Predicted Versus Actual Costs for High-Cost Incidents, Overall and by Component

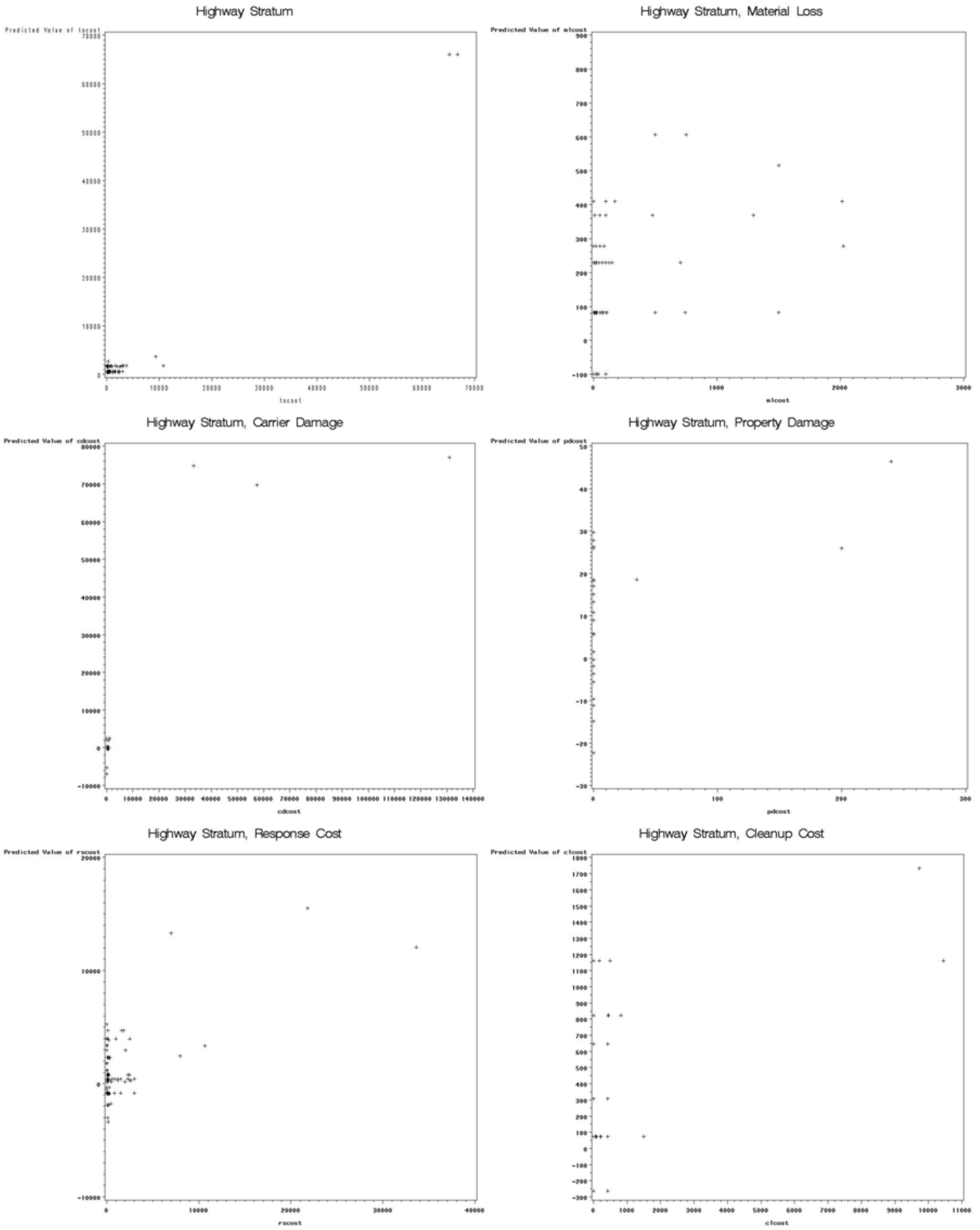


Figure A-2. Predicted Versus Actual Costs for Highway Incidents, Overall and by Component

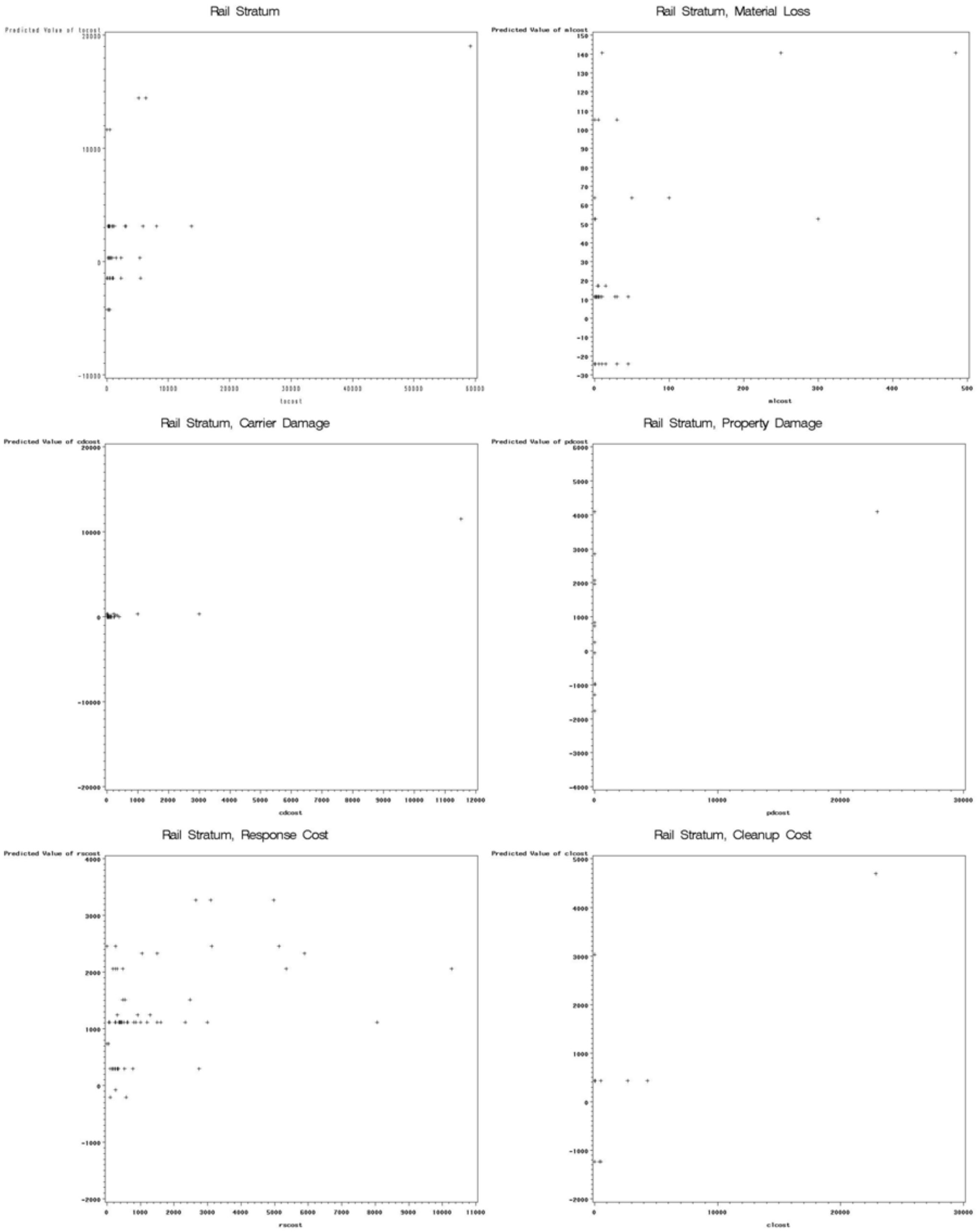


Figure A-3. Predicted Versus Actual Costs for Rail Incidents, Overall and by Component

APPENDIX B

SERIOUS INCIDENTS

As discussed in Chapter 2, this study was designed to estimate total costs and costs in four strata defined by high-cost incidents and mode. The high-cost stratum was developed as a mechanism to reduce the variability in the costs by mode by separating out incidents that were identified as "unusual events." While there is some overlap between the definition of a high-cost incident and a "Serious" incident as defined by the Department of Transportation, there are some differences as well. PHMSA's 2002 definition of a serious incident states that serious incidents involve at least one of the following conditions:

- A fatality or major injury caused by the release of a hazardous material;
- The evacuation of 25 or more persons as a result of release of a hazardous material or exposure to fire;
- A release or exposure to fire which results in the closure of a major transportation artery;
- The alteration of an aircraft flight plan or operation;
- The release of radioactive materials from Type B packaging;
- The release of over 11.9 gallons or 88.2 pounds of a severe marine pollutant; or
- The release of a bulk quantity (over 119 gallons or 882 pounds) of a hazardous material.

There were a total of 83 serious incidents in the selected sample, of which 49 had complete cost information. The 83 incidents were distributed among the sampling strata as follows:

- 78 were in the high-cost stratum,
- 4 were in the highway stratum, and
- 1 was in the rail stratum.

In all four highway incidents and the rail incident, the reason for being a serious incident but not a high-cost incident was that the quantity of release exceeded the threshold. In the HMIRS database, there were 456 serious incidents between November 1, 2002, and October 31, 2003. Of these:

- 126 were in the high-cost stratum;
- 273 were in the highway stratum (268 because of volume; 8 because of evacuations; 3 had both);
- 45 were in the rail stratum (all 45 because of volume; 1 also met the evacuation criteria);
- 10 were in the air stratum (all 10 because of flight plan alteration);
- 2 were in the water stratum (1 because of volume, 1 because of evacuations).

The same methodology described in the main body of this report was used to analyze the serious incidents. In short, sample estimates were calculated separately for each cost component and for total costs (stratified sampling estimate). Regression curves were also fit for each cost component and for total costs (stratified regression estimate). The results for serious incidents are consistent with those found for the high-cost incidents. In particular, the sum of the total reported costs in the HMIRS database for serious incidents is $34,184,053; whereas, the stratified sampling estimate is roughly four times (3.7 times) higher at $127,112,312 and the stratified regression estimate is roughly twice as high (1.9 times) at $64,274,670. Table B-1 provides details, including estimated confidence intervals, for both the stratified sampling estimates and the stratified regression estimates.

Table B-1. Estimates for Average and Total Incident Costs (in Dollars) for Serious Incidents, Overall and by Component

Stratum	Material Loss	Carrier Damage	Property Damage	Response Cost	Cleanup Cost	Total
Number in HMIRS	456	456	456	456	456	456
Number in Sample	55	53	56	53	55	49
Average Incident Cost ($)	8,002	63,508	15,180	81,580	89,176	278,755
Std Error of Incident Cost ($)	11,741	86,999	67,204	111,158	169,922	344,835
Reported Total Cost						34,184,053
Stratified Sampling Estimates						
Total Costs ($)	3,649,003	28,959,827	6,922,134	37,200,412	13,006,858	127,112,312
Std Error of Total Cost ($)	676,998	5,122,821	3,835,430	6,545,529	9,797,698	21,222,342
Lower 95% Confidence Bound for Total ($)	2,322,087	18,919,098	≤ 0	24,371,175	≤ 0	85,516,522
Upper 95% Confidence Bound for Total ($)	4,975,919	39,000,556	14,439,577	50,029,649	32,210,346	168,708,102
Stratified Regression Estimates						
Total Costs ($)	1,944,879	6,524,127	6,193,124	21,915,425	16,697,116	64,274,670
Std Error of Total Cost ($)	335,815	3,265,680	2,366,823	4,773,401	7,566,178	12,015,491
Lower 95% Confidence Bound for Total ($)	1,288,682	123,394	1,554,151	12,559,559	1,867,407	40,724,308
Upper 95% Confidence Bound for Total ($)	2,603,076	12,924,860	10,832,097	31,271,291	31,526,825	87,825,032

APPENDIX C

SUPPORTING DOCUMENTATION FOR DATA COLLECTION

Section 2.2, Data Collection, discussed the approach used to obtain cost information from the reporting entities as well as from third parties, including shippers, cleanup companies, and response agencies. This Appendix contains three tools used by the data collection team:

1. A telephone script for their initial call to the carrier that reported the incident; accompanied by a shorter script for leaving a voice message.

2. Answers to Frequently Asked Questions (FAQs).

3. Sample e-mail text to send to the contact, providing a detailed list of the requested information, including subcategories for each of the five major cost areas.

Hazmat Incident Cost Project – Telephone Script

Mr./Ms. XXXXX,

I am calling on behalf of the U.S. Department of Transportation's Office of Hazardous Materials Safety. My name is *XXXXX* and I work for Battelle, which in under contract to DOT. The DOT is studying hazardous materials transportation incident costs and has selected one of *company*'s incidents for this study. I'm calling you since you signed the incident report submitted to DOT.

For this work, DOT is trying to determine total costs for hazmat transportation incidents and to develop methods for estimating these costs for similar incidents in the future. This improved cost information will allow the DOT to make better decisions about the cost-benefit of future regulatory initiatives and possibly reduce the need to collect these data on the 5800 incident report form.

We are collecting detailed information on incident costs broken down into several categories. What I'd like to do is get your e-mail address and send you a list of the specific cost information that we are looking for, since you probably don't have it all right in front of you right now. Then you can collect the information and call me back so we can discuss it.

The specific incident that we are interested in is one that occurred on *mm/dd/yyyy* in *city, state* and that involved *type of operation* of *material*.

Do you have any questions about our study? *(see the FAQs)*

Are you the appropriate person to provide cost information for this/these incident(s)? May I please get your e-mail address so I can send you more detail on the specific information we are looking for? *(if no e-mail address, obtain current fax number or mailing address)*

Thanks. When do you think you will be able to collect the information for that/those incident(s)? *(they may not be able to estimate this without seeing the specific information you are looking for – if they can't commit, indicate that you'll call them back in a week or so to touch base with them)*

I look forward to speaking with you soon *(or "on XXXXX")* and really appreciate your assistance on this project.

Hazmat Incident Cost Project – Voice Mail Script

Mr./Ms. XXXXX,

I am calling on behalf of the U.S. Department of Transportation's Office of Hazardous Materials Safety concerning a hazmat transportation incident involving *company*. I'm calling you since you signed the incident report submitted to DOT. My name is *XXXXX* and I work for Battelle, which in under contract to DOT. My number is 202-646-xxxx and I hope to hear from you soon. Thanks.

Hazmat Incident Cost Project – FAQs

1. Do I have to provide the information you are requesting?
   No. At this point, we are asking for voluntary participation. However, it is important to our statistical analysis that we obtain data on most of the incidents that have been selected. This project will give the DOT better information on the costs of hazmat incidents and allow them to make better decisions about the cost-benefit of future regulatory initiatives and possibly reduce the need to collect these data on the 5800 incident report form.
   [If pressed, we do have regulatory authority—49 CFR 171.21—but are not exercising it.]

2. When do you need this information?
   We would like it as soon as you can gather it; however, we understand that you have a business to run and understand if there are some delays in compiling all the requested data. I ask that you please keep us informed about any expected delays you may have in responding.
   [As we get closer to the end of the project, we might give a date when the data must be in.]

3. What if my boss needs proof or documentation that you are who you say you are?
   I can provide you with supporting documentation from the U.S. DOT and a contact name there if you have any additional questions. *[Make sure we have a correct e-mail, fax, or mailing address.]*
   [We have a copy of the first few pages of our contract with DOT]
   [Our DOT project manager: Ron DiGregorio, 202-366-0644, ronald.digregorio@dot.gov]

4. How was this incident selected?
   We developed a statistical approach to sample four to five hundred incidents, based on mode, total reported costs, fatalities and injuries, and other characteristics. This approach will allow us to apply the results of our analysis to all 15,000 incidents that occur each year.

5. Will I be asked for data on other incidents?
   We have identified all incidents in the current sample that belong to your company; however, there will be a total of two samples taken from all incidents that occurred from November 2002 through October 2003. The initial sample will cover the first eight months and the second will cover the next four months. It is possible that one of your incidents will be selected in a later sample.
   [This can be modified depending on which sample you are working from.]

Hazmat Incident Cost Project – Sample E-mail

Mr./Ms. XXXXX,

Thank you for your assistance on this project. As I mentioned on the phone, Battelle (www.battelle.org/transportation) is under contract with the U.S. DOT's Office of Hazardous Materials Safety to study incident costs for a sample of recent incidents, one of which was the *company* incident in *city, state* on *mm/dd/yyyy*, which involved *type of operation* of *material*. This incident was assigned a DOT incident report number of *XXXXX*. *(if more than one incident was sampled for the same company, modify this paragraph to refer to all of them)*

For this work, Battelle is performing additional research and follow up for statistically selected hazardous materials transportation incidents to determine total incident costs and to develop methods for estimating these costs for similar incidents in the future. This improved cost information will allow the DOT to make better decisions about the cost-benefit of future regulatory initiatives and possibly reduce the need to collect these data on the 5800.1 incident report form.

We are also interested in learning about the decisions that carriers make in completing the incident cost fields to better understand the data that are currently reported and to develop appropriate guidance for completing these fields in the future.

If you wish, I can provide some additional documentation to substantiate our contract with DOT to collect this data on their behalf. I'd prefer to go over these questions with you on the phone when you have a chance to collect the pertinent information from your files and accounting systems. Relevant information may be spread across your safety, operations, legal, insurance, or accounting departments.

Hopefully, you will be able to provide much of the information we need for this incident; however, we are happy to contact other parties directly to obtain additional information if necessary. This could include the shipper, cleanup contractors, and response companies or agencies as well as others within your own company. Note that not all of the cost items listed below will pertain to this specific incident, but are included for completeness.

Modify the items below to be specific to each incident or group of incidents.

Material loss

- The reported amount for product loss was $XXXX. How did you determine the "value" of the lost product that was shipped? Was any remaining product accepted and paid for by the consignee? If not, then the total cost of the shipped material would be appropriate here. I assume that this information is not listed on your shipping invoices.

Carrier damage

- Costs to repair or replace the damaged cargo tank or tractor
- Costs resulting from damage to other cargo from the hazardous material
- Damages paid to third parties, such as insurance claims or lawsuits (unless directly related to another category, such as property damage)
- Towing/removal costs for the cargo tank or other company-owned vehicles
- Costs for investigations, reporting, documentation, and communications related to the incident (this would include labor hours, travel, equipment rental, etc. – please estimate these if your do not specifically tracked them). Actual costs to manage the response are separate and included in ‘response cost’ below

Property damage

- Repair and replacement costs of other vehicles
- Repair and replacement costs to buildings and other fixed facilities
- Restoration of open land not included as response or cleanup
- Repair and replacement costs of other non-hazardous materials or packaging
- Insurance claims paid for property damage

Response cost

- Costs incurred by local emergency response from police and fire departments and local or regional emergency response teams – if these costs were not passed on to you, I would like to contact them directly to get an estimate of their costs
- Costs incurred by you to an emergency response services vendor
- Costs incurred directly by you – this would include the costs to dispatch company employees to the scene to manage the incident (labor hours, travel, equipment costs, etc.) and is distinct from the costs listed under 'carrier damage' above

Cleanup cost

- Disposal costs (e.g., collecting, transporting, and ultimately disposing of all material collected during the response phase)
- Remediation costs (e.g.; excavation; disposal and replacement of contaminated soil; pumping, treatment, and re-injection of contaminated groundwater; and absorption and disposal of hazardous materials released into surface water)

Where appropriate, I would like to include labor and travel costs for all parties. Also, it is possible that some costs are ongoing or a final cost accounting has not been made. For these costs, please indicate when you think final cost information will be available and try to provide an estimate of the total expected costs in each of the five areas listed above.

Are there any other costs you can think of that are not included above?

Please call me at the number below when you have collected the requested information. I look forward to speaking with you soon. Thanks again for your assistance.

Signature