U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

NHTSA

DOT HS 812 795

September 2019

Crash Report Sampling System: Imputation

DISCLAIMER

Suggested APA Format Citation:

Herbert, G. C. (2019, September). *Crash Report Sampling System: Imputation* (Report No. DOT HS 812 795). Washington, DC: National Highway Traffic Safety Administration.

1. Report No. DOT HS 812 795	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Crash Report Sampling System: Imputation		5. Report Date September 2019
		6. Performing Organization Code NSA-210
7. Author Gabrielle C. Herbert		8. Performing Organization Report No.
9. Performing Organization Name and Address Mathematical Analysis Division, National Center for Statistics and Analysis National Highway Traffic Safety Administration 1200 New Jersey Avenue SE Washington, DC 20590		10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address National Highway Traffic Safety Administration 1200 New Jersey Avenue SE Washington, DC 20590		13. Type of Report and Period Covered NHTSA Technical Report 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

This report documents the imputation procedures applied to a select number of key data elements in Crash Report Sampling System (CRSS). In 2016 the CRSS replaced the National Automotive Sampling System General Estimates System (NASS GES) to modernize NHTSA's data collection systems. The overall imputation methodology has not changed between CRSS and NASS GES. The imputation procedure has been applied to 2016 and 2017 CRSS data. This is the first time the CRSS imputation process has been documented.

17. Key Words	18. Distribution Statement
Crash Report Sampling System, CRSS, CRSS imputation, sequential regression multivariate imputation, univariate imputation, NASS GES.	No restrictions. This document is available to the public through the National Technical Information Service, www.ntis.gov.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	30	

Table of Contents

1. Executive Summary

To provide vital information on motor vehicle traffic crashes, the National Highway Traffic Safety Administration annually publishes nationally representative estimates of police-reported motor vehicle traffic crashes and their characteristics. From 1988 to 2015 NHTSA created national estimates using data from the National Automotive Sampling System General Estimates System (NASS GES), which sampled police crash reports from police jurisdictions across the United States. In 2016 NHTSA replaced NASS GES with the Crash Report Sampling System (CRSS), which is representative of the same crash population, to modernize the data collection system.

From sampled police crash reports, analysts code information into approximately 120 different data elements. In some instances item nonresponse occurs due to missing or a lack of detailed information from the sampled police crash report. To resolve this issue, NHTSA's National Center for Statistics and Analysis (NCSA) employs a statistical technique called "imputation" to assign values to data items that are unknown or not reported.

We apply imputation techniques to CRSS data because they can reduce potential bias and allow replicability and comparability of published information across multiple years. Currently, NCSA applies sequential regression multivariate imputation (SRMI) and univariate imputation methods to data attributes that are unknown or not reported for a select number of key data elements. These imputation techniques remain the same as those that were applied to NASS GES data.

In total, 27 key data elements are imputed within the CRSS accident, vehicle, and person data files. Each imputed data element uses multiple iterations of imputation through the SRMI or univariate imputation methods. After each round of imputation, checks are applied to verify imputed data are accurate and consistent with other data elements in the police-reported crash.

In summary, CRSS encounters data item nonresponse like NASS GES. To deal with this ongoing issue, NHTSA continues to apply imputation techniques to 27 key data elements from three CRSS data files. We apply these techniques to CRSS to minimize potential bias and generate reproducible and consistent results across multiple years.

2. Introduction

The National Highway Traffic Safety Administration has established multiple data collection systems to gather important information on motor vehicle crashes. In the 1970s NHTSA began collecting data on police-reported motor vehicle traffic crashes through the National Automotive Sampling System. NASS eventually separated into the NASS General Estimates System and NASS Crashworthiness Data System (NASS CDS) during the 1980s. To modernize NHTSA's data collection systems, the Crash Report Sampling System and Crash Investigation Sampling System (CISS) replaced NASS GES and NASS CDS, respectively. This report focuses on the statistical imputation procedures applicable to CRSS. Refer to Zhang, Noh, Subramanian, & Chen (in press) for more information on CISS.

The primary purpose of CRSS is to create national estimates of police-reported motor vehicle traffic crashes as well as their characteristics. Since the collection of all police-reported motor vehicle traffic crashes is currently unobtainable, approximately 50,000 police crash reports are selected using a three-stage probability sampling method. To produce unbiased national estimates, weights are assigned to each sampled police crash report. Due to nonresponding sampling units (i.e., unit nonresponse), nonresponse adjustments as well as calibration factors are applied to calculated design weights accordingly. Refer to Zhang, Noh, Subramanian, & Chen (2019, May) for more information regarding the CRSS sample design and weighting.

CRSS collects information only from police crash reports. They provide a multitude of information about traffic crashes through fill-in data items as well as crash narratives and diagrams (refer to Appendix A for an example of a police crash report). Information provided on each police crash report is coded into a database. During the coding process, analysts code certain CRSS data elements as "not reported" or "unknown" because information is missing, or due to a lack of detailed information in the narrative or crash diagram. Information from police crash reports may also be unavailable for some data elements because of illegible data items, no data section or block to fill in information on the police crash report, or inconsistency of reported information. When data items like those collected from police crash reports are missing, unknown, or not reported, it is referred to as "item nonresponse." To handle this issue, NHTSA applies statistical imputation, which replaces an unknown or not reported response with a value, to CRSS data (Brick & Kalton, 1996). Even though some information is missing, unknown, or not reported from the police crash report, the rate of unknown or not reported data in CRSS is relatively low for almost all imputed key data elements.

In the following chapters, we discuss the history of NASS GES imputation, the purpose of CRSS imputation imputed CRSS data elements, the imputation methodologies used in CRSS, and the CRSS imputation procedure and data validation. Finally, we will discuss the limitations of imputation. This is the first time the CRSS imputation process has been documented.

3. Overview and History of NASS GES Imputation

Imputation is a widely used statistical technique to deal with item nonresponse in surveys and data collection systems. Over the years, NHTSA's NCSA has applied different imputation techniques to unknown or not reported police-reported traffic crash data characteristics. The following chapter describes these various statistical techniques.

To produce annual motor vehicle traffic safety reports of crash indicators, NHTSA statisticians applied the summary level imputation technique to unknown and not reported NASS data. For this technique, unknown and not reported data was proportionately distributed to the reported data attributes for each published table. Summary level imputation was relatively simple and potentially reduced bias within published tables. It also required manual application of the imputation technique to every summary table and data checks of consistency from table to table, thus adding additional production time and delayed the release of tables. Also, released data files did not provide users with imputed information. Unless users followed the imputation procedure exactly, replication of results was very difficult (Shelton, 1993).

NHTSA statisticians continued to apply the summary level imputation method to NASS GES. Due to extended production time and the manual nature of summary level imputation, the univariate imputation method replaced summary level imputation in the 1990s. Shelton (1993) describes the univariate imputation method as "The process of randomly substituting known values for one variable for unknown values for the same variable." Unlike summary level imputation, each unknown or not reported record received a non-missing value from the univariate imputation method. This gave NHTSA the ability to provide data users with imputed data on public files. Even though univariate imputation was more manageable and less time consuming than summary level imputation, univariate imputation still had its challenges. It only works well for a variable with a low missing item rate or one with little association with other variables on the data files.

To overcome the shortcomings of the univariate imputation method, NHTSA statisticians began to apply hot-deck imputation to data elements where univariate imputation was deemed unsuitable. Unlike univariate imputation, where imputation is solely based on the data element, "hot-deck" imputation assigns unknown or not reported data to known values based on records with similar data attributes. To apply hot-deck imputation, a set of data elements determined to be highly correlated with the imputed variable, also known as classification variables, are established prior to imputation using the chi-square test of independence for each data element. For a given imputed data element, the associated classification variables are used to create an n-dimensional table where n is the number of classification variables. In a table cell of the classification variables, records are randomly sorted. Each record with an unknown or not reported data attribute is assigned a value from a non-missing record with the same data attributes for each classification variable. If an unknown item is unable to be assigned, then a classification variable is removed from the n-dimensional table, thus creating an (n-1) dimensional table. This step is repeated for a set number of iterations where a classification variable is removed during each iteration. If hot-deck failed to impute, univariate imputation was used to impute the remaining unknown or not reported items. See Shelton (1993) for more information on NASS GES imputation techniques.

For the 2010 NASS GES, NHTSA determined the current imputation methodology should be evaluated and potentially updated. After review, the SRMI method ultimately replaced hot-deck imputation. In addition, data elements imputed using the univariate imputation method became limited due to the introduction of the SRMI method. SRMI uses statistical stepwise regression models to assign values to unknown or not reported data. It was implemented due to drawbacks with univariate and hot-deck imputation methods. When data elements were introduced or removed from the NASS GES file, or data element attributes changed, hot-deck imputation required manual maintenance of classification variables to determine if the relationship between variables remained the same. Also, as data is imputed, more information becomes available, which can lead to changes in the relationship between two data elements but hot-deck imputation only considered the relationship between the imputed data element and classification variables at the beginning of the imputation process. Finally, NHTSA had not revised the NASS GES imputation methodology in more than 15 years, which was not employing the most modern imputation techniques. SRMI provided a new and modern approach to imputation, while providing an automated selection of different statistical models and related predictors throughout the process for an imputed data element. All key data elements identified for imputation used the SRMI method except for *Body Type,* where only univariate imputation was applied. If imputed data from SRMI was inconsistent with other data element attributes or statistical issues arose with SRMI, the univariate imputation method was still applied to the remaining unknown or not reported data. NASS GES continued to apply SRMI and univariate imputation methods to data elements until 2015 (NCSA, 2011).

4. Purpose of CRSS Imputation

For many data collection systems, item nonresponse can produce major challenges for data users including reduced sample size for analysis, biased results, and the possibility of inconsistent data from user to user. This chapter details how imputation can resolve these CRSS data challenges.

When unknown or not reported data is present, it may cause biased results if ignored (Brick & Kalton, 1996). Imputation can potentially reduce the bias that results from unknown or not reported data. For instance, if we assume there are 100 people, 10 people are injured, 60 people are not injured, and 30 people have unknown/not reported injury severity. Users may interpret the results as 10 people had an injury. If we know 15 people with unknown injury severity actually had injuries, then the true number of people with injuries is 25. The original conclusion is misleading and biased. Shelton (1993) states, "Imputation is one way of trying to reduce this bias. By making intelligent "guesses" about the unknowns, the bias may very well be reduced."

To minimize misleading conclusions as seen in the example above, NHTSA uses imputed data from CRSS to produce estimates of motor vehicle traffic crash characteristics. For example, Table 72 in the NHTSA (2017) Traffic Safety Facts 2015 report uses the imputed data elements, *Maximum Injury Severity in a Crash*, *Initial Point of Impact*, and *Body Type* to provide a summary table of Vehicle Occupants Injured by Initial Point of Impact and Vehicle Type for injured occupants. By using imputed data, the bias in the summary table may be reduced.

Because of the complete case analysis (i.e., only using records with non-missing values), statistical programs or users can handle unknown or not reported data in a variety of ways allowing for possible reduction in sample size or inaccurate results. Brick and Kalton state, "Survey analysis is multivariate in nature and low item response rates for several items together may result in a sizeable proportion of records with missing data for one or more of the items involved in a particular analysis. Thus, for example, a substantial proportion of records may be dropped when a multiple regression analysis with many variables is restricted to records with complete data for all variables involved." Imputation eliminates the need to discard records with unknown or not reported data items for computations. NHTSA provides data users with access to imputed values on CRSS to allow users to carry out statistical computations on complete data elements. Also, the availability of imputed values gives data users the ability to produce consistent and comparable results with NHTSA and other data users (Brick & Kalton, 1996).

5. Imputed CRSS Data Elements

CRSS data files contain approximately 120 unique data elements, which provide detailed information on characteristics of motor vehicle traffic crashes. Any of the coded data elements can potentially have unknown or not reported data attributes but only a few key data elements are imputed. Key imputation variables have remained the same since 2010 NASS GES imputation. In total, 27 data elements are imputed, which include 12 data elements from the accident file, 9 from the vehicle file, and 6 from the person file. NHTSA selected these variables for imputation because they are used to produce key summary data tables[1] that provide overall crash indicators of overall surveillance of the traffic safety issues. Also, data users frequently use these variables. Table 1 lists the imputed data elements. NHTSA derives some data elements from other imputed data elements on the person file. Asterisks (*) denote derived data elements in Table 1.

Table 1: Imputed Data Elements on Crash Report Sampling System

Data File	Imputed Data Elements
ACCIDENT	Alcohol Involved in Crash*, Atmospheric Conditions, Crash Date – Day of the Week, Crash Time – Hour, Crash Time – Minute, First Harmful Event, Light Condition, Manner of Collision, Maximum Injury Severity in Crash*, Number of Injured in Crash*, Relation to Junction Within Interchange Area, Relation to Junction – Specific Location
VEHICLE	Areas of Impact – Initial Contact Point, Body Type, Driver Drinking in Vehicle*, Hit and Run, Number of Injured in Vehicle*, Maximum Injury Severity in Vehicle*, Most Harmful Event, Model Year, Pre-Movement Prior to Critical Event
PERSON	Age, Ejection, Injury Severity, Police – Reported Alcohol Involvement, Seat Position, Sex

To preserve the original coded information, NHTSA creates new data elements for imputed data. All imputed data elements have "_IM" as a suffix. For example, the imputed data element name for AGE is AGE_IM. These new data elements include the original coded data, when known/reported, as well as the imputed data (i.e., there are no unknown or not reported data attributes). Appendix B details the original data element and corresponding imputed data element name along with the SAS label.

[1] See Traffic Safety Facts 2015 (NHTSA, 2017) for tables using the 27 imputed data elements. At the time of this report, the 2016 and 2017 Traffic Safety Facts annual reports with CRSS data had not been publicly released but the imputed NASS GES data elements apply the same imputation methodology as CRSS.

6. CRSS Imputation Methodology

From 2010 to 2015, NHTSA applied the SRMI method to NASS GES. NASS GES also implemented the univariate imputation method when deemed necessary (NCSA, 2011). CRSS continues to use these two imputation methodologies. The following chapter further describes the two imputation methods used in CRSS.

6.1 Sequential Regression Multivariate Imputation

Raghunathan, Lepkowski, Van Hoewyk, and Solenberger (2001) created the SRMI method to handle complex survey data structure such as data element restrictions and boundaries. The Survey Research Center of the Institute for Social Research at the University of Michigan developed the SAS-callable software, Imputation and Variance Estimation Software (IVEware)[2] to implement the SRMI method (Raghunathan, Solenberger, & Van Hoewyk, 2011).

SRMI relies on coded data or previously imputed data to produce fully conditional explicit regression models for each imputed data element. Based on the imputed data element data type (which is specified by the data user), SRMI can either build a normal linear regression model for continuous data, a logistic regression model for binary data, a generalized logit regression model for categorical data, a Poisson loglinear regression model for count data, or a two-stage regression model using a logistic and normal linear regression models for mixed data. Since most key data elements on CRSS are categorical, the SRMI method produces generalized logit regression models for those data elements. SRMI uses the posterior predictive distribution of statistical models to impute values for each unknown or not reported data item. SRMI method is briefly described below[3] (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001).

1. Determine data elements with unknown/not reported values, $Y_1 \ldots Y_k$ where Y_1 is the data element with the least number of unknown/not reported values and Y_k is the data element with the most number of unknown/not reported values.
2. Based on the data type of Y_1, regress Y_1 on X using the appropriate regression model where X is a matrix of non-missing predictors with an assumed flat prior distribution.
3. Draw a single value from posterior predictive distribution for each unknown/not reported value in Y_1 based on the associated regression model.
4. Regress Y_2 on X and Y_1 and draw values from posterior predictive distribution for each unknown/not reported value.
5. Regress the remaining data elements $Y_3 \ldots Y_k$ on X, Y_1, $Y_2 \ldots Y_{k-1}$ and select values for unknown/not reported values from the corresponding posterior predictive distribution.
6. Continue to regress Y_k on X, $Y_1, \ldots, Y_k$ (i.e., all predictors – including imputed data elements – except the dependent variable) then draw from the posterior predictive distribution and replace previously imputed values until imputed values are stable[4] or for a defined number of regression cycle iterations.[5]

[2] University of Michigan. (2012). IVEware: Imputation and Variance Estimation Software (version 0.2), www.src.isr.umich.edu/software/iveware-downloads/version-2/

[3] See a more detailed explanation of the sequential regression multivariate imputation method in Raghunathan, Lepkowski, Van Hoewyk, & Solenberger. (2001)

[4] Imputation values are considered stable when values meet defined convergence criteria.

[5] IVEware recommends 10 cycles for most imputations (Raghunathan, Solenberger, & Van Hoewyk, 2011).

As mentioned previously, SRMI creates statistical regression models based on all predictors or data elements with non-missing values. Due to the sample size and number of data elements on the CRSS data files, computational time of the SRMI method becomes extremely extensive. To reduce computational time, each model is created using a stepwise regression model. NHTSA statisticians define the model selection criteria prior to imputation.[6] Stepwise regression models follow similar steps as described above but imputed data elements are regressed on selected predictors based on an IVEware algorithm (Raghunathan, Solenberger, & Van Hoewyk, 2011).

6.2 Univariate Imputation

In some instances, imputed data is inconsistent with non-missing data based on consistency checks. Also, IVEware can fail when defined convergence criteria is not met. For these cases, univariate imputation, or simple random sample with replacement, is applied to the remaining inconsistent or unimputed records. NHTSA implements the univariate imputation method with SAS 9.4. The univariate imputation method is detailed below (Zhang, Noh, Subramanian, & Chen, 2019).

1. Identify and separate the unknown or not reported values and non-missing values (i.e., observed values and imputed values from SRMI or univariate imputation) into two groups for the imputed data element.
2. Select a random value (with replacement) from the group of non-missing values and assign the random value to an unknown or not reported record.
3. Repeat the previous step for the remaining unknown or not reported cases in the data element.

Unlike SRMI, where imputation is dependent on other data elements, univariate imputation ignores any correlation to other data elements. If a data element has a low rate of unknown or not reported values or little correlation with other data elements, then the univariate imputation is a plausible technique. Since most data elements have very few unknown values after SRMI, univariate imputation is an acceptable imputation method for the remaining cases. The univariate imputation preserves the non-missing distribution of a data element after imputation (Shelton, 1993).

[6] The minimum marginal r-squared required for a predictor to be included in the model is set to 0.01. The maximum number of predictors in a model 15.

7. CRSS Imputation Procedure

CRSS imputes unknown or not reported values for data elements from the accident, vehicle, and person files using single imputation (i.e., a single plausible value is assigned to an unknown or not reported value). CRSS imputation uses SRMI, univariate imputation and derivation (logical imputation) to impute data elements.[7] This chapter details the CRSS imputation procedure: data preparation, imputation of each data file, derivation of data elements from imputed data, and consistency checks and data element checks. Finally, we give an example of data validation.

7.1 Data Preparation

Prior to the application of imputation techniques, NHTSA statisticians review unknown/not reported item rates of imputed data elements and compare results to prior years. Appendix D contains the unknown/not reported rates for all imputed data elements on the 2017 CRSS data files. Unknown/not reported rates for most key data elements are relatively low with a few exceptions. *Relation to Junction Within Interchange Area* (21.1%) and *Police-Reported Alcohol Involvement* (29.8%), have higher rates of unknown/not reported values compared to other data elements. *Alcohol Involved in Crash* (14.4%), and *Driver Drinking in Vehicle* (9.4%) also have high rates of unknown/not reported values, which are derived from *Police-Reported Alcohol Involvement.* Most data element unknown/not reported rates remain consistent from year to year. In rare instances, shifts to unknown/not reported rates could occur. Usually these shifts are due to updates to police-reporting or data coding procedures.

Coding analysts code all CRSS data elements with a specific value including unknown or not reported attributes (i.e., no blank/missing items). Even though known and unknown/not reported data elements have different numeric values, IVEware and univariate imputation SAS programs do not consider any coded information as a candidate for imputation. Therefore, any unknown or not reported data attributes for imputed data elements must be "flagged" (i.e., set to missing) for imputation. Additionally, possible covariates may have numerous data attributes which can cause issues like excessive computational time or too many model predictors in IVEware. These elements are collapsed into smaller categories based on classifications defined by NHTSA.

7.2 Data File Imputation

Imputation of the three data files occurs in four phases. They are based on the hierarchical structure of the CRSS data files. The phases of imputation are as follows:

[7] Refer to Appendix C for results of the 2017 CRSS imputation cases by imputation methodology.

Imputation Phase	Data File	Imputed Data Elements
1	ACCIDENT	Atmospheric Conditions, Crash Date – Day of the Week, Crash Time – Hour, Crash Time – Minute, First Harmful Event, Light Condition, Manner of Collision, Relation to Junction Within Interchange Area, Relation to Junction – Specific Location
2	VEHICLE	Body Type
3	VEHICLE	Areas of Impact – Initial Contact Point, Hit and Run, Most Harmful Event, Model Year, Pre-Movement Prior to Critical Event
4	PERSON	Age, Ejection, Injury Severity, Police – Reported Alcohol Involvement, Seat Position, Sex

After each phase of imputation, the coded and imputed data elements are merged to the subsequent data file. This allows all data elements, including imputed and coded data, to be used as possible predictors for the SRMI method or for consistency checks. For instance, prior to imputation of the data elements on the person file, all the data elements from the accident and vehicle file are merged with the data elements on the person file.

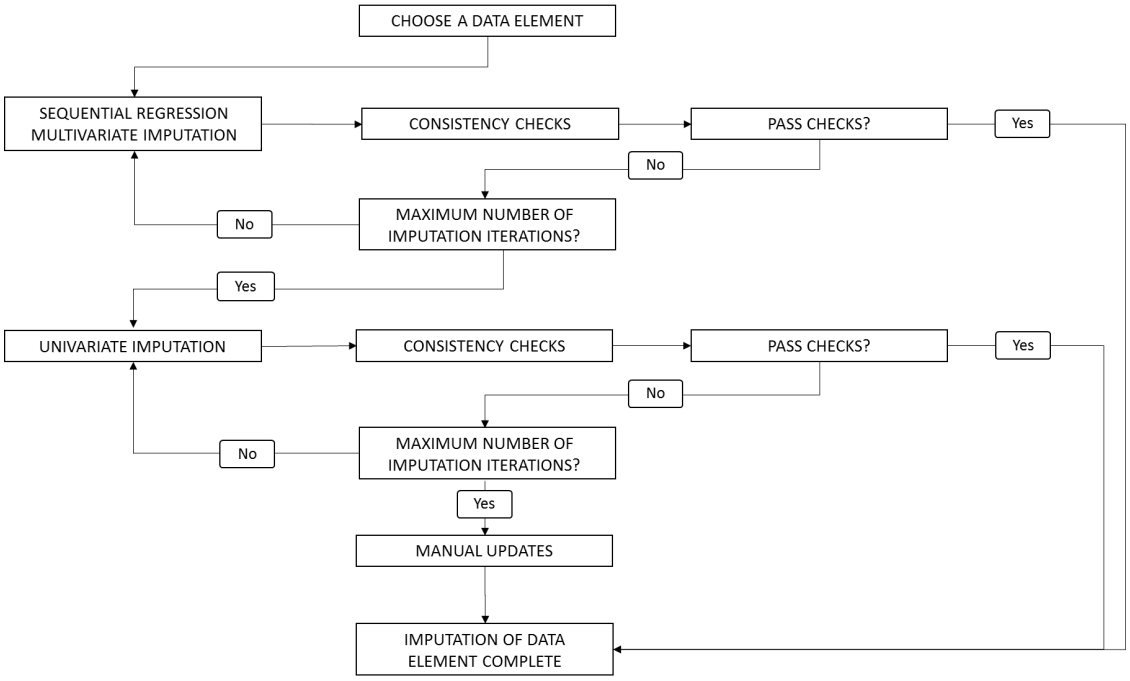
During Phase 1, Phase 3, and Phase 4, data elements are imputed using the SRMI[8] method through IVEware and the univariate imputation method. First, IVEware produces stepwise regression models to assign values to unknown or not reported CRSS data. In some instances, a two-stage imputation is applied when a data element has many data attributes or levels, NHTSA has predefined categories for a data element, the data element naturally separates into categories, or a data element has multiple unknown levels for some categories. Two-stage imputation initially aggregates the data element into specific groups or categories. For example, *Age* is aggregated into categories based on predefined NHTSA groups. IVEware assigns unknown or not reported data attributes to a category. After the category is assigned, a value within the given category (i.e., de-aggregated category) is assigned using IVEware. To continue the example, if a record is imputed to the category of "10- to 15-year-olds" then the record would be assigned an age from 10 to 15. Two-stage imputation is used to impute *Age* and *Seating Position* on the person file.

After values have been assigned, the imputed data are checked for consistency with other coded or imputed data. For more information on consistency checks see Chapter 7.4. If there are any inconsistencies, these cases are identified for another iteration or round of imputation using the SRMI method. Additionally, IVEware may fail when convergence criteria is not met. Failure to impute cases are identified for another iteration of imputation using the SRMI method as well. The SRMI method along with consistency checks are executed an additional four iterations. After the multiple iterations of the SRMI method and consistency checks, imputed data may still be inconsistent with non-missing data or IVEware may still fail. The univariate imputation

[8] See Appendix E for the selected regression models type for data elements from Phase 1, Phase 3, and Phase 4.

method is applied to any lingering inconsistent or fail to impute cases. Consistency between imputed data and non-missing data is reevaluated. Again, inconsistencies may still be an issue and are identified for another iteration of univariate imputation. Like the SRMI method, univariate imputation and consistency checks are executed an additional four times. Upon completion, there may be still a few cases with inconsistencies. Based on the FARS/CRSS Coding Manual (NHTSA, 2018), CRSS Analytical User's Manual (NHTSA, 2019) and motor vehicle traffic crash knowledge, NHTSA statisticians review the cases to determine the best consistent and plausible attribute for the remaining cases. Figure 1 details the imputation process for Phase 1, Phase 3, and Phase 4.

Figure 1: CRSS Imputation Procedure using SRMI and Univariate Imputation Method



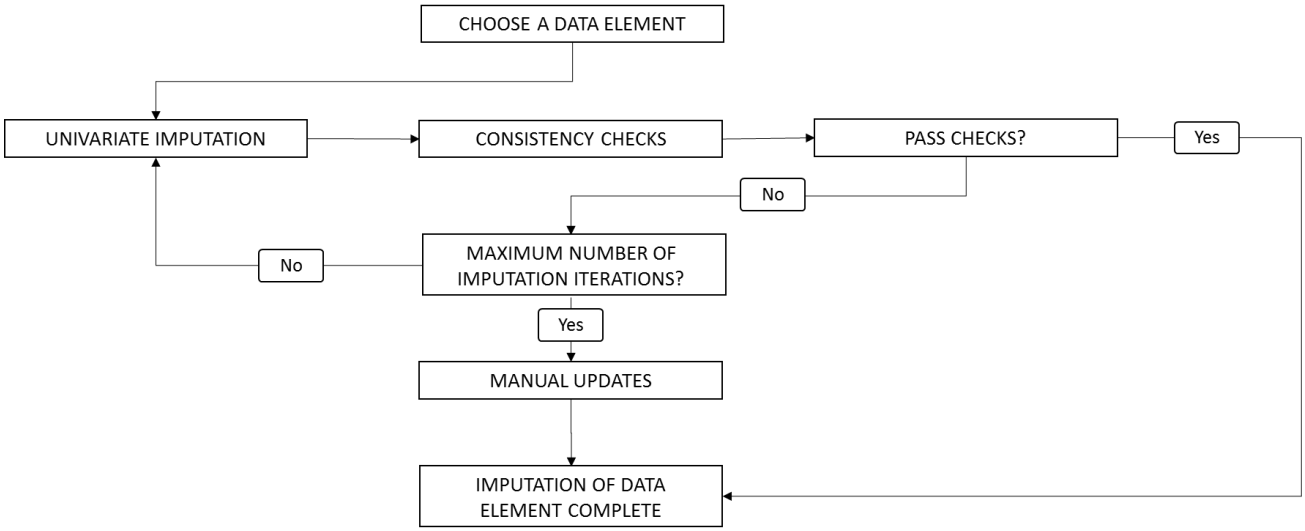
As stated before, *Body Type* is imputed separately from other imputed data elements on the vehicle file (Phase 2). *Body Type* is imputed using only the univariate imputation method to preserve the distribution of the data element prior imputation and allow the data element to be used for consistency checks or as a possible predictor for imputation of other vehicle level data elements. Body type contains four levels of unknown data attributes, "Unknown Light Vehicle Type," "Unknown Truck Type," "Not Reported," and "Unknown Body Type". The four unknown data attributes are imputed in the following three separate steps.

1. Unknown Light Vehicle Type
2. Unknown Truck Type
3. Not Reported and Unknown Body Type

First, the univariate imputation method is applied to unknown light vehicles and each case is checked for consistency. It is important to note the data is subset to only include light vehicles.

This process is repeated for ten iterations. After imputation and consistency checks, some cases may still be inconsistent. As previously noted, NHTSA statisticians review the cases and manually update the case to a consistent and plausible *Body Type*. Figure 2 details the imputation process for Phase 2. The same process is applied to unknown trucks and not reported/unknown body type. For unknown truck imputation, the data is subset to only include trucks. For general not reported/unknown body type, the all body types are included for imputation.

Figure 2: CRSS Imputation Procedure for Univariate Imputation Method



7.3 Data Element Derivation

After Phase 4, the imputed data as well as the original coded data from *Injury Severity* and *Police-Reported Alcohol Involvement* are used to derive/logically impute six variables on the accident and vehicle files. These derived variables include *Alcohol Involved in Crash*, *Number of Injured in Crash*, *Maximum Injury Severity in Crash*, *Driver Drinking in Vehicle*, *Number Injured in Vehicle* and *Maximum Injury Severity in Vehicle*. Table 2 details the derived data element and the associated person file data elements used to derive the data element.

Table 2: Derivation of Accident and Vehicle Data File Elements

Data File	Derived Variable	Person Data File Elements Used to Derive Variables
ACCIDENT	Alcohol Involved in Crash	• Person Type • Police-Reported Alcohol Involvement
ACCIDENT	Number of Injured in Crash	• Injury Severity
ACCIDENT	Maximum Injury in Crash	• Injury Severity
VEHICLE	Driver Drinking in Vehicle	• Person Type • Police-Reported Alcohol Involvement
VEHICLE	Number of Injured in Vehicle	• Injury Severity
VEHICLE	Maximum Injury Severity in Vehicle	• Injury Severity

Driver Drinking in Vehicle and *Alcohol Involved in Crash* are derived from the data elements, *Police-Reported Alcohol Involvement* and *Person Type*, from the person file. If *Driver Drinking in Vehicle* is unknown, and a role of a person involved in a crash is considered a "Driver of a Motor Vehicle in Transport," then the imputed value of *Driver Drinking in Vehicle* would be assigned the corresponding value of the *Police-Reported Alcohol Involvement*. *Alcohol Involved in Crash* is derived similarly to *Driver Drinking in Vehicle*, but the role of a person involved in the crash includes drivers as well as non-motorists.

The remaining four derived variables, *Number of Injured in Crash*, *Maximum Injury Severity in Crash*, *Number of Injured in Vehicle* and *Maximum Injury Severity in Vehicle* are calculated from the data element *Injury Severity* on the person file. The *Number of Injured in Crash* and the *Number of Injured in Vehicle* use reported and imputed injury severity to calculate the total number of injured people in a crash or vehicle, respectively. The number of injured data elements include individuals with fatal injuries, suspected serious injuries, suspected minor injuries, possible injuries as well as injuries with unknown severity. The *Maximum Injury Severity in a Crash* and the *Maximum Injury Severity in Vehicle* are derived based on a specified ranking of *Injury Severity*. The ranking of Injury Severity is as follows: "Fatal," "Suspected Serious Injury," "Suspected Minor Injury," "Possible Injury," "Injured, Severity Unknown," "No Apparent Injury," "Died Prior to Crash" (NHTSA, 2019). *Injury Severity* is ranked for each person in a crash and vehicle. The most severe injury (coded or imputed) determines the value of *Maximum Injury Severity in Crash* and *Maximum Injury Severity in Vehicle*.

It is important to note the derived variables may not have to same value has the original reported value (even though they may not be coded as unknown or not reported). For example, if the *Maximum Injury in a Vehicle* is reported as "Possible Injury" and a vehicle occupant has "Unknown/Not Reported' injury severity and *Injury Severity* is imputed to "Suspected Minor Injury" for that vehicle occupant, then the *Maximum Injury in a Vehicle* would be "Suspected Minor Injury."

7.4 Imputation Consistency Checks and Data Element Checks

CRSS is a complex data collection system where many of the data elements are related or rely on other data elements. To deal with the complexity, consistency checks are applied to each data element from the police crash report. Consistency checks are detailed in the Fatality Analysis Reporting System/Crash Report Sampling System Coding and Validation Manual. See NHTSA (2018) for the 2017 consistency checks for CRSS. For example, if the crash *Month* is between May and September then the *Atmospheric Conditions* (WEATHER) should not equal "Sleet or Hail," "Snow," "Blowing Snow," or "Freezing Rain or Drizzle" (NHTSA, 2018). During and after the coding process, an intensive edit checking process occurs based on consistency checks. If there are any inconsistencies, coding analysts make necessary updates. This allows coded information including unknown or not reported information to be accurate prior to imputation. Once data is imputed, there is a chance the imputed data is inconsistent with the original coded data or other imputed data. To resolve these issues, consistency check programs evaluate and determine if there are any issues with the imputed data. As previously mentioned if there are inconsistencies, NHTSA statisticians re-impute or manually update the record. It is important to note; consistency checks are only completed on imputed data. Each check ensures the data is accurate and consistent.

In addition to consistency checks, data validation is evaluated for the entire data element. After SRMI, univariate imputation and derivation of data elements, the distribution prior to application of the imputation technique (i.e., pre-imputation) is reviewed and compared to the data element distribution after imputation technique (i.e., post-imputation). If the percentage of imputed values for a data element's attributes (especially for data attributes with rare occurrences) seems unreasonable, the data is assessed and issues are resolved by another iteration of imputation or manual updates. Additionally, after all data is imputed, the original data element and the imputed data element are compared to verify there were no differences in coded/reported values between the two data elements. This was not done for derived data elements because they are subject to change. Finally, all values were verified to have valid values (e.g., within boundaries, non-negative values).

7.5 Data Validation Example

In this section, we discuss an example to demonstrate the validity of imputed CRSS data. The example is based on the data element, *Injury Severity* (INJ_SEV), from the 2017 PERSON file. Currently, *Injury Severity* only has 3.46 percent (n= 4,802) of the data unknown/not reported. It is imputed using the SRMI method. For this example, a simple random sample of 20 percent (n=26,823) is set to unknown for *Injury Severity*. The imputation methods described above are

applied to the artificial unknown data. From the SRMI method, the predictor data elements for *Injury Severity* are *Vehicle Removal* (TOWED) and *Transported to First Treatment Facility* (HOSPITAL). These are the same predictors selected for *Injury Severity* during the 2017 CRSS imputation process. When the results from the artificial imputed data and the observed values are compared, about two-thirds of the records matched. Additionally, we collapsed the artificial unknown data into the following categories: "No Injury," "Injury," "Fatal Injury" to evaluate the results. It is collapsed because most evaluations of injury severity consider an injured person as any injury type that is not fatal (i.e., possible injury, suspected minor injury, suspected serious injury, and injured, severity unknown). The collapsed observed data compared to the artificial data resulted in an almost 80 percent (78.8%) match between the three categories. Table 3 and Table 4 compares the distribution of the observed data and artificial imputed data for the randomly selected records set to unknown.

Table 3. Unweighted *Injury Severity* Distribution for Observed and Artificial Data

	Observed Data	Artificial Imputed Data
No Apparent Injury	68.3%	68.3%
Possible Injury	15.9%	15.9%
Suspected Minor Injury	9.5%	9.0%
Suspected Serious Injury	5.2%	5.4%
Fatal Injury	0.7%	1.1%
Injured, Severity Unknown	0.4%	0.4%
Died Prior to Crash	0.0%	-

Source: CRSS 2017

Table 4. Unweighted Collapsed *Injury Severity* Distribution for Observed and Artificial Data

	Observed Data	Artificial Imputed Data
No Injury	68.3%	68.3%
Injury	30.9%	30.6%
Fatal Injury	0.7%	1.1%

Source: CRSS 2017

The unweighted distribution of the observed compared to the artificial imputed data is relatively similar for Table 3 and Table 4. Based on Table 4, about 68 percent of people are not injured, about 31 percent are injured and about 1 percent are fatally injured for the observed and artificial imputed data. Table 5 and Table 6 compare the total observed, total observed and imputed, and total observed and artificial imputed *Injury Severity* data.

Table 5. Unweighted *Injury Severity* Distribution Comparison

	Observed Data	Observed & Imputed Data*	Observed & Artificial Imputed Data
No Apparent Injury (O)	68.4%	68.7%	68.7%
Possible Injury (C)	15.8%	15.8%	15.8%
Suspected Minor Injury (B)	9.2%	9.0%	8.9%
Suspected Serious Injury (A)	5.4%	5.3%	5.3%
Fatal Injury (K)	0.8%	0.8%	0.9%
Injured, Severity Unknown	0.4%	0.4%	0.4%
Died Prior to Crash*	0.0%	0.0%	0.0%

Source: CRSS 2017
* "Observed & Imputed Data" is INJ_SEV on the 2017 PERSON File.

Table 6. Unweighted Collapsed *Injury Severity* Distribution Comparison

	Observed Data	Observed & Imputed Data*	Observed & Artificial Imputed Data
No Injury	68.4%	68.7%	68.7%
Injury	30.8%	30.5%	30.4%
Fatal Injury	0.8%	0.8%	0.9%

Source: CRSS 2017
* "Observed & Imputed Data" is INJSEV_IM on the 2017 PERSON file.

The unweighted distribution between the data is almost the same for Table 5 and Table 6. With 20 percent of the data unknown compared to 3.5 percent unknown, there is less information available for creation of stepwise regression models, yet the distribution between the observed data and the artificial imputed data remain consistent for each table. Thus, verifying the data from our imputation methodology is valid.

8. Limitations

While the data produced from imputation is useful in many aspects, it has some limitations. The general objective of CRSS imputation is to provide values to estimate key statistics for NHTSA's Traffic Safety Facts. Only essential data elements are imputed from CRSS, leaving data users to handle unknown or not reported items for all other data elements (NCSA, 2011). While some data users may use an imputation methodology like the ones described in this report, other data users may use a different approach which may cause inconsistent results from user to user.

As noted earlier, the CRSS imputation methodologies rely on information from imputed data elements as well as non-imputed data elements to assign the most plausible values to unknown or not report cases. Because of this, imputed values are highly dependent on the quality of the coded data. If coded information is incorrect, it could potentially affect the imputed values. Additionally, since only a portion of data elements are imputed, item nonresponse of other data elements can possibly influence the quality of the imputation process.

CRSS implements single imputation which replaces a single missing or in the case of CRSS unknown or not reported values with a non-missing value. While single imputation has advantages like a complete dataset (or data elements), potential reduction in bias and consistency between data users, single imputation assumes the imputed value is an observed value (i.e., no uncertainty in the value). This assumption can lead to an underestimation of the variance and standard errors. Underestimation of variance and standard errors can be resolved through multiple imputation. Multiple imputation produces multiple sets of plausible datasets, which can take into account the uncertainty of imputed values. Since NHTSA provides the public with imputed values for each data element, multiple sets of plausible imputed values may cause issues with application of results for some users (Rubin, 1988).

References

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, Vol. 5*, 215-238.

National Center for Statistics and Analysis (2011). *Imputation in the 2010 NASS GES.* (Unpublished research note). Washington, DC: National Highway Traffic Safety Administration.

National Highway Traffic Safety Administration. (2018, October). *2017 Fatality Analysis Reporting System/Crash Report Sampling System coding and validation manual* (Report No. DOT HS 812 559). Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812559

National Highway Traffic Safety Administration. (2019, April). *Crash Report Sampling System analytical user's manual 2016 -2017* (Report No. DOT HS 812 702). Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812702

National Highway Traffic Safety Administration. (2017). *Traffic safety facts 2015: A Compilation of motor vehicle crash data from the Fatality Analysis Reporting System and the General Estimates System* (Report No. DOT HS 812 384). Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812384

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001, June). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, Vol. 27*, No. 1, 85-95.

Raghunathan, T.E., Solenberger, P.W., Van Hoewyk, J. (2011, September). *IVEware: Imputation and Variance Estimation Software Version 0.2 Users Guide (Supplement)*. Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

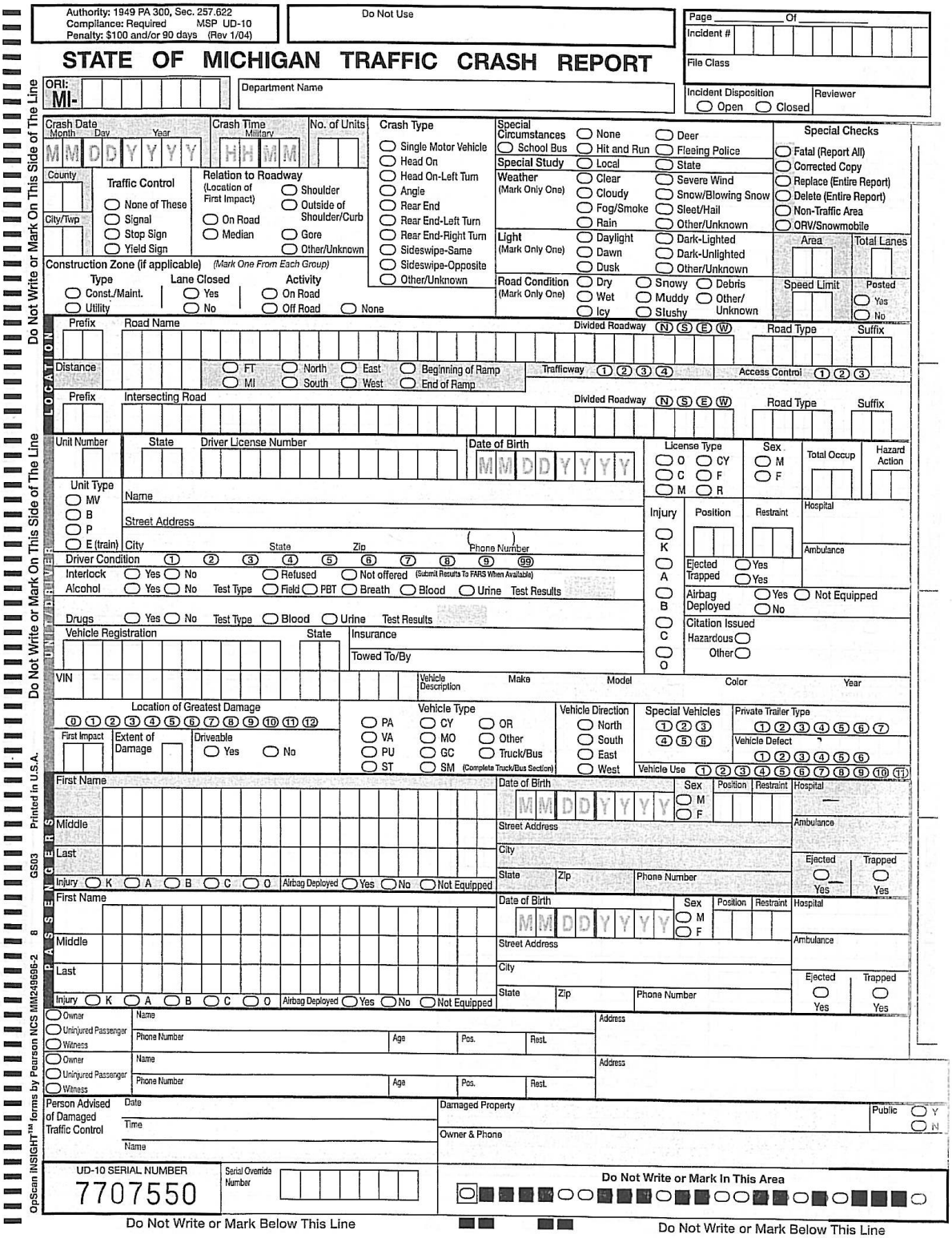
Rubin, D. B. (1988). *An overview of multiple imputation.* In Proceedings of the Survey Research Section, American Statistical Association, Alexandria, VA.

Shelton, T. S. T. (1993, June). *Imputation in the NASS General Estimates System* (Report No. DOT HS 807 985). Washington, DC: National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/807985

Zhang, F., Noh, E. Y., Subramanian, R., & Chen, C.-L. (in press). *Crash Investigation Sampling System: Sample design and weighting.* Washington, DC: National Highway Traffic Safety Administration.

Zhang, F., Noh E. Y., Subramanian R., Chen C.-L. (2019, May) *Crash Report Sampling System: Sample design and weighting.* (Report No. DOT HS 812 706). Washington, DC: National Highway Traffic Safety Administration.

Appendix A: Example of Police Crash Report



Authority: 1949 PA 300, Sec. 257.622
Compliance: Required MSP UD-10
Penalty: $100 and/or 90 days (Rev 1/04)

Do Not Use

Page ____ Of ____

Incident #

File Class

Incident Disposition ○ Open ○ Closed

Reviewer

STATE OF MICHIGAN TRAFFIC CRASH REPORT

ORI: MI-

Department Name

Crash Date Month Day Year (MMDDYYYY)

Crash Time Military (HHMM)

No. of Units

Crash Type
○ Single Motor Vehicle
○ Head On
○ Head On-Left Turn
○ Angle
○ Rear End
○ Rear End-Left Turn
○ Rear End-Right Turn
○ Sideswipe-Same
○ Sideswipe-Opposite
○ Other/Unknown

Special Circumstances ○ None ○ Deer ○ School Bus ○ Hit and Run ○ Fleeing Police
Special Study ○ Local ○ State
Weather (Mark Only One) ○ Clear ○ Severe Wind ○ Cloudy ○ Snow/Blowing Snow ○ Fog/Smoke ○ Sleet/Hail ○ Rain ○ Other/Unknown
Light (Mark Only One) ○ Daylight ○ Dark-Lighted ○ Dawn ○ Dark-Unlighted ○ Dusk ○ Other/Unknown
Road Condition (Mark Only One) ○ Dry ○ Snowy ○ Debris ○ Wet ○ Muddy ○ Other/Unknown ○ Icy ○ Slushy

Special Checks
○ Fatal (Report All)
○ Corrected Copy
○ Replace (Entire Report)
○ Delete (Entire Report)
○ Non-Traffic Area
○ ORV/Snowmobile

Area Total Lanes

Speed Limit Posted ○ Yes ○ No

County

City/Twp

Traffic Control
○ None of These
○ Signal
○ Stop Sign
○ Yield Sign

Relation to Roadway (Location of First Impact)
○ On Road ○ Shoulder ○ Median ○ Outside of Shoulder/Curb ○ Gore ○ Other/Unknown

Construction Zone (if applicable) (Mark One From Each Group)
Type: ○ Const./Maint. ○ Utility
Lane Closed: ○ Yes ○ No
Activity: ○ On Road ○ Off Road ○ None

LOCATION

Prefix Road Name Divided Roadway (N)(S)(E)(W) Road Type Suffix

Distance ○ FT ○ MI ○ North ○ South ○ East ○ West ○ Beginning of Ramp ○ End of Ramp Trafficway (1)(2)(3)(4) Access Control (1)(2)(3)

Prefix Intersecting Road Divided Roadway (N)(S)(E)(W) Road Type Suffix

UNIT/DRIVER

Unit Number State Driver License Number Date of Birth (MMDDYYYY)

License Type ○ O ○ CY ○ C ○ F ○ M ○ R
Sex ○ M ○ F
Total Occup
Hazard Action

Unit Type ○ MV ○ B ○ P ○ E (train)

Name

Street Address

City State Zip Phone Number

Injury ○ K ○ A ○ B ○ C ○ O
Position
Restraint
Hospital
Ambulance

Driver Condition (1)(2)(3)(4)(5)(6)(7)(8)(9)(99)
Interlock ○ Yes ○ No ○ Refused ○ Not offered (Submit Results To FARS When Available)
Alcohol ○ Yes ○ No Test Type ○ Field ○ PBT ○ Breath ○ Blood ○ Urine Test Results
Drugs ○ Yes ○ No Test Type ○ Blood ○ Urine Test Results

Ejected ○ Yes Trapped ○ Yes
Airbag Deployed ○ Yes ○ No ○ Not Equipped
Citation Issued
Hazardous ○
Other ○

Vehicle Registration State
Insurance
Towed To/By

VIN
Vehicle Description Make Model Color Year

Location of Greatest Damage (0)(1)(2)(3)(4)(5)(6)(7)(8)(9)(10)(11)(12)
First Impact
Extent of Damage
Driveable ○ Yes ○ No

Vehicle Type ○ PA ○ CY ○ OR ○ VA ○ MO ○ Other ○ PU ○ GC ○ Truck/Bus ○ ST ○ SM (Complete Truck/Bus Section)
Vehicle Direction ○ North ○ South ○ East ○ West
Special Vehicles (1)(2)(3)(4)(5)(6)
Private Trailer Type (1)(2)(3)(4)(5)(6)(7)
Vehicle Defect (1)(2)(3)(4)(5)(6)
Vehicle Use (1)(2)(3)(4)(5)(6)(7)(8)(9)(10)(11)

PASSENGERS

First Name Middle Last
Date of Birth (MMDDYYYY) Sex ○ M ○ F Position Restraint Hospital
Street Address
City State Zip Phone Number
Ejected ○ Yes Trapped ○ Yes
Injury ○ K ○ A ○ B ○ C ○ O Airbag Deployed ○ Yes ○ No ○ Not Equipped

First Name Middle Last
Date of Birth (MMDDYYYY) Sex ○ M ○ F Position Restraint Hospital
Street Address
City State Zip Phone Number
Ejected ○ Yes Trapped ○ Yes
Injury ○ K ○ A ○ B ○ C ○ O Airbag Deployed ○ Yes ○ No ○ Not Equipped

○ Owner ○ Uninjured Passenger ○ Witness
Name Phone Number Age Pos. Rest. Address

○ Owner ○ Uninjured Passenger ○ Witness
Name Phone Number Age Pos. Rest. Address

Person Advised of Damaged Traffic Control
Date
Time
Name

Damaged Property
Owner & Phone
Public ○ Y ○ N

UD-10 SERIAL NUMBER 7707550

Serial Override Number

Do Not Write or Mark In This Area

Do Not Write or Mark Below This Line

Do Not Write or Mark Below This Line

Do Not Write or Mark On This Side of The Line

Do Not Write or Mark On This Side of The Line

OpScan INSIGHT™ forms by Pearson NCS MM249695-2 8 GS03 Printed in U.S.A.

Appendix B: SAS Names for Imputed Values

File Name	Original Variable Name	Imputed Variable Name	SAS Label
ACCIDENT	ALCOHOL	ALCHL_IM	Imputed Drinking in Crash
ACCIDENT	DAY_WEEK	WKDY_IM	Imputed Day of the Week
ACCIDENT	HARM_EV	EVENT1_IM	Imputed First Harmful Event
ACCIDENT	HOUR	HOUR_IM	Imputed Hour
ACCIDENT	LGT_COND	LGTCON_IM	Imputed Lgt Condition
ACCIDENT	MINUTE	MINUTE_IM	Imputed Minute
ACCIDENT	MAN_COLL	MANCOL_IM	Imputed Manner of Collision
ACCIDENT	MAX_SEV	MAXSEV_IM	Imputed Maximum Injury Severity
ACCIDENT	NUM_INJ	NO_INJ_IM	Imputed Number Injured in Crash
ACCIDENT	RELJCT1	RELJCT1_IM	Relation to Junction – Within Interchange Area
ACCIDENT	RELJCT2	RELJCT2_IM	Imputed Relation to Junction – Junction
ACCIDENT	WEATHER	WEATHR_IM	Imputed Weather Condition
VEHICLE	IMPACT1	IMPACT1_IM	Imputed Area of Impact-Initial
VEHICLE	BODY_TYP	BDYTYP_IM	Imputed Body Type
VEHICLE	VEH_ALCH	V_ALCH_IM	Imputed Driver Drinking in Vehicle
VEHICLE	HIT_RUN	HITRUN_IM	Imputed Hit and Run
VEHICLE	MAX_VSEV	MXVSEV_IM	Imputed Maximum Injury in Vehicle
VEHICLE	MOD_YEAR	MDLYR_IM	Imputed Model Year
VEHICLE	P_CRASH1	PCRASH1_IM	Imputed Vehicle P_Crash1
VEHICLE	M_HARM	VEVENT_IM	Imputed Most Harmful Event
VEHICLE	NUM_INJV	NUMINJ_IM	Imputed Number Injured in Vehicle
PERSON	AGE	AGE_IM	Imputed Age
PERSON	EJECTION	EJECT_IM	Imputed Ejection
PERSON	INJ_SEV	INJSEV_IM	Imputed Injury Severity
PERSON	DRINKING	PERALCH_IM	Imputed Police Rep. Alcohol Inv.
PERSON	SEAT_POS	SEAT_IM	Imputed Seating Position
PERSON	SEX	SEX_IM	Imputed Sex

Appendix C: 2017 CRSS Cases Imputed by Imputation Methodology[9]

Data File	Variable	Description	Number of Cases Imputed by SRMI	Number of Cases Imputed by Univariate Imputation	Number of Cases Manually Updated	Number of Cases Derived from Imputed Variables	Total Number of Cases Imputed
ACCIDENT	ALCOHOL	Alcohol Involved in Crash	-	-	-	7,915	7,915
ACCIDENT	DAY_WEEK	Crash Date (Day of Week)	-	-	-	-	-
ACCIDENT	HARM_EV	First Harmful Event	36	-	1	-	37
ACCIDENT	HOUR	Crash Time (Hour)	152	-	-	-	152
ACCIDENT	LGT_COND	Light Condition	428	-	-	-	428
ACCIDENT	MINUTE	Crash Time (Minute)	152	-	-	-	152
ACCIDENT	MAN_COLL	Manner of Collision	266	-	1	-	267
ACCIDENT	MAX_SEV	Maximum Injury Severity in Crash	-	-	-	989	989
ACCIDENT	NUM_INJ	Number of Injured in Crash	-	-	-	989	989
ACCIDENT	RELJCT1	Relation to Junction - Within Interchange Area	11,574	-	-	-	11,574
ACCIDENT	RELJCT2	Relation to Junction - Specific Location	1,134	7	1	-	1,142
ACCIDENT	WEATHER	Atmospheric Conditions	735	1,735	-	-	2,470
VEHICLE	IMPACT1	Area of Impact - Initial	2,450	2	-	-	2,452

[9] The totals are based on 2017 CRSS data. These values are subject to change from year to year.

Data File	Variable	Description	Number of Cases Imputed by SRMI	Number of Cases Imputed by Univariate Imputation	Number of Cases Manually Updated	Number of Cases Derived from Imputed Variables	Total Number of Cases Imputed
		Contact Point					
VEHICLE	BODY_TYP	Body Type	-	2,737	12	-	2,749
VEHICLE	VEH_ALCH	Driver Drinking in Vehicle	-	-	-	9,206	9,206
VEHICLE	HIT_RUN	Hit and Run	7	1	-	-	8
VEHICLE	MAX_VSEV	Maximum Injury Severity in Vehicle	-	-	-	4,079	4,079
VEHICLE	MOD_YEAR	Model Year	3,702	-	-	-	3,702
VEHICLE	P_CRASH1	Pre-Movement Prior to Critical Event	2,017	12	1	-	2,030
VEHICLE	M_HARM	Most Harmful Event	35	-	-	-	35
VEHICLE	NUM_INJV	Number of Injured in Vehicle	-	-	-	4,079	4,079
PERSON	AGE	Age	4,976	4,620	-	-	9,596
PERSON	EJECTION	Ejection	8,896	1	2	-	8,899
PERSON	INJ_SEV	Injury Severity	4,802	-	-	-	4,802
PERSON	DRINKING	Police-Reported Alcohol Involvement	41,357	-	1	-	41,358
PERSON	SEAT_POS	Seat Position	2,319	6	9	-	2,334
PERSON	SEX	Sex	5,819	-	-	-	5,819

Appendix D: 2017 CRSS Rates of Unknown/Not Reported Values[10]

File Name	Variable Name	Description	Unknown/Not Reported Rate
ACCIDENT	ALCOHOL	Alcohol Involved in Crash	14.40%
ACCIDENT	DAY_WEEK	Crash Date (Day of Week)	0.00%
ACCIDENT	HARM_EV	First Harmful Event	0.07%
ACCIDENT	HOUR	Crash Time (Hour)	0.28%
ACCIDENT	LGT_COND	Light Condition	0.78%
ACCIDENT	MINUTE	Crash Time (Minute)	0.28%
ACCIDENT	MAN_COLL	Manner of Collision	0.49%
ACCIDENT	MAX_SEV	Maximum Injury Severity in Crash	1.80%
ACCIDENT	NUM_INJ	Number of Injured in Crash	1.80%
ACCIDENT	RELJCT1	Relation to Junction – Within Interchange Area	21.06%
ACCIDENT	RELJCT2	Relation to Junction – Specific Location	2.08%
ACCIDENT	WEATHER	Atmospheric Condition	4.49%
VEHICLE	IMPACT1	Area of Impact – Initial Contact Point	2.51%
VEHICLE	BODY_TYP	Body Type	2.82%
VEHICLE	VEH_ALCH	Driver Drinking in Vehicle	9.43%
VEHICLE	HIT_RUN	Hit and Run	0.01%
VEHICLE	MAX_VSEV	Maximum Injury Severity in Vehicle	4.18%
VEHICLE	MOD_YEAR	Model Year	3.79%
VEHICLE	P_CRASH1	Pre-Event Movement	2.08%
VEHICLE	M_HARM	Most Harmful Event	0.04%
VEHICLE	NUM_INJV	Number Injured in Vehicle	4.18%
PERSON	AGE	Age	6.91%
PERSON	EJECTION	Ejection	6.41%
PERSON	INJ_SEV	Injury Severity	3.46%
PERSON	DRINKING	Police-Reported Alcohol Involvement	29.77%
PERSON	SEAT_POS	Seating Position	1.68%
PERSON	SEX	Sex	4.19%

[10] The percentages are based on 2017 CRSS data. The values are subject to change from year to year.

Appendix E: Regression Model Type for CRSS Data Elements Imputed by SRMI

Regression Model Type	Imputed Variable
Normal Linear Model	• Crash Time (Minute) • Model Year
Logistic Model	• Relation to Junction - Within Interchange Area • Hit and Run • Police-Reported Alcohol Involvement • Sex
Generalized Logit Model	• First Harmful Event • Crash Time (Hour) • Light Condition • Manner of Collision • Relation to Junction - Specific Location • Atmospheric Conditions • Area of Impact - Initial Contact Point • Pre-Movement Prior to Critical Event • Most Harmful Event • Age • Ejection • Injury Severity • Seat Position

DOT HS 812 795
September 2019

U.S. Department
of Transportation

**National Highway
Traffic Safety
Administration**

14322-090919-v3