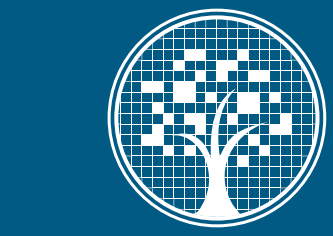


HEALTHCARE COST AND UTILIZATION PROJECT

HCUP Methods Series

Missing Data Methods for the NIS and the SID

Report # 2015-01

Agency for Healthcare Research and Quality

U.S. Department of Health and Human Services
Agency for Healthcare Research and Quality

Contact Information:
Healthcare Cost and Utilization Project (HCUP)
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
http://www.hcup-us.ahrq.gov

For Technical Assistance with HCUP Products:

Email: hcup@ahrq.gov

or

Phone: 1-866-290-HCUP

Recommended Citation: Houchens R. *Missing Data Methods for the NIS and the SID*. 2015. HCUP Methods Series Report # 2015-01 ONLINE. January 22, 2015. U.S. Agency for Healthcare Research and Quality. Available: http://www.hcup-us.ahrq.gov/reports/methods/methods.jsp.

TABLE OF CONTENTS

EXECUTIVE SUMMARY 1

1. Introduction 3

2. Types of Missing Data 4

2.1 Missing Completely at Random 4
2.2 Missing at Random 5
2.3 Missing Not at Random 5

3. Missing Data in the NIS and SID Files 6

3.1 Missing Data in the NIS 2012 6
3.2 Missing Data in the Michigan SID 2012 9

4. Missing Data Methods 12

4.1 Case Deletion Method 12
4.2 Single Imputation Methods 12
4.3 Weighting Methods for Finite Population Statistics 13
4.4 Likelihood-Based Methods 13
4.5 Multiple Imputation Methods 14
4.5.1 Common Imputation Models 15
4.5.1.1 Regression Imputation 16
4.5.1.2 Predictive Mean Matching Imputation 16
4.5.1.3 Propensity Score Imputation 17
4.5.1.4 Logistic Regression Imputation 17
4.5.1.5 Discriminant Function Imputation 17
4.5.2 Multivariate Imputation 17
4.5.2.1 Markov Chain Monte Carlo Imputation 18
4.5.2.2 Full Conditional Specification Imputation 18
4.5.2.3 Monotone Imputation 19

5. Missing Data Software 19

5.1 SAS 20
5.2 Stata 20
5.3 R 21

6. Examples 21

6.1 SID Examples 22
6.2 NIS Examples 33

7. Recommendations 44

8. References 46

Appendix A: SAS Code for Analysis of Complicated Diabetes, Michigan SID 2012
Appendix B: SAS Code for Analysis of Complicated Diabetes, NIS 2012

INDEX OF TABLES

Table 1. Percentage of missing values for selected data elements, NIS 2012 8

Table 2. Percentage of missing values for selected data elements, Michigan SID 2012...10

Table 3. Missing data patterns for complicated diabetes, Michigan SID 2012.....24

Table 4. Missing data patterns for complicated diabetes, group means, Michigan SID 2012.....24

Table 5. Pooled versus original regression coefficients for complicated diabetes, Michigan SID 2012 31

Table 6. Missing data patterns for complicated diabetes, NIS 2012.....34

Table 7. Missing data patterns for complicated diabetes, group means, NIS 2012.....35

Table 8. Percentage by primary payer and imputation for complicated diabetes, NIS 2012 39

Table 9. Pooled versus original regression coefficients for complicated diabetes, NIS 2012.....42

INDEX OF FIGURES

Figure 1. Distribution of observed and imputed log(charges) 28
Figure 2. Effect of age on total charges for complicated diabetes, original 32
Figure 3. Distribution of observed and imputed age for complicated diabetes, NIS 2012 37
Figure 4. Distribution of observed and imputed log(LOS+1) 38
Figure 5. Comparison of observed and imputed distribution of log(charges) 40
Figure 6. Multiplicative effect of age on total charges for complicated diabetes, ... 43

EXECUTIVE SUMMARY

This report is intended to guide users and raise their awareness about the need to address missing data in the Healthcare Cost and Utilization Project (HCUP) using the National (Nationwide) Inpatient Sample (NIS), the State Inpatient Databases (SID), and other HCUP data. HCUP is a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality (AHRQ).

HCUP data represent rich sources of information for health services researchers. The National Inpatient Sample (NIS) annually provides a nationally representative sample of hospital discharge records that can be used to study relationships among hospital outcomes, discharge characteristics, and hospital characteristics at the national, regional, and census division levels. For States that participate in the HCUP Central Distributor, each State Inpatient Database provides a near-census of annual hospital discharge data for that State.

Other HCUP databases include the Kids' Inpatient Database (KID) for studies of children's health, the Nationwide Emergency Department Sample (NEDS) for the study of emergency department utilization, the State Ambulatory Surgery and Services Database (SASD) for the study of ambulatory surgery, and the State Emergency Department Databases (SEDD), which contain a near-census of emergency department visits that can be used to study State-specific emergency department utilization for States that choose to release their data through the HCUP Central Distributor.

All of these databases have data elements with missing values for a portion of their records. In fact, missing values are inevitable in most research databases. Typically, users simply discard records with missing values for key data elements and then generate estimates solely on the basis of records without missing values. This behavior is encouraged because it is the default method for handling missing values in most statistical software. However, unless the missing values occur in a purely random fashion, this approach can lead to biased estimates. Also, by eliminating cases with missing values, potentially valuable information contained in the nonmissing data is discarded.

For example, suppose that (unknown to the user) a data source set hospital charges to a missing value for discharges from all rural hospitals. Then if charges differed between rural and nonrural hospitals, an estimate of overall mean charges based solely on nonmissing charges would be biased. In the literature, this approach to missing values is called *listwise deletion* or *complete case analysis*.

Imputation methods fill in the missing data with plausible values allowing all of the data to be used in the analysis. It can help overcome any biases inherent in complete case analysis, which is valid only when data are *missing completely at random* (MCAR), meaning that the probability of a missing value is the same for all cases. Unfortunately, it is usually impossible to know whether data are MCAR. Therefore, whenever data contain missing values it is good practice to at least try imputation to test whether the results are sensitive to the missing values. Moreover, there is no established threshold on the rate of missing values below which

imputation is clearly unnecessary. Under some conditions, even a very low missing value rate can have an adverse effect on statistical estimates.

Our main recommendation is that typical HCUP data users should use a missing data technique called *multiple imputation*, which is widely accepted and available in all of the major statistical packages. This technique imputes M multiple plausible values for each missing value that occurs in an analytic file. Then a separate estimate for each parameter of interest is generated from M complete data sets, each with a different set of imputed values. Finally, by a process called *Rubin's Rules*, the M estimates are pooled to form a single estimate for each parameter and its standard error.

In this report, we give an overview of missing data methods and work through some examples using the NIS 2012 and the Michigan SID 2012. This report is meant as an introduction to missing data methods and to missing values in HCUP data. Armed with this information, users are encouraged to complete their education on missing data methods by consulting references that are cited throughout this report.

1. INTRODUCTION

This report is *not* a general tutorial on missing data methods. Several excellent books and articles on missing data methods, many of which are cited in this report, explain the theory and application of missing data methods, often illustrated with real-world missing data problems. Instead, this report is meant to guide users and to raise their awareness about the need to address missing data in the Healthcare Cost and Utilization Project (HCUP) using the National (Nationwide) Inpatient Sample (NIS), the State Inpatient Databases (SID), and other HCUP data. HCUP is a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality (AHRQ).

The NIS is a database of U.S. hospital discharge data designed to inform policy decisions regarding health and health care at the national level, at the census region level, and starting with the NIS 2012, at the census division level. Through NIS data, researchers can make inferences about national trends in health care utilization, access, cost, quality, and outcomes. The NIS is the largest all-payer inpatient care database that is publicly available in the United States and has been made publicly available since the 1988 data year.

For any given year, the NIS contains a sample of between 7 and 8 million hospital discharges from U.S. community hospitals, representing a 20 percent sample of discharges nationally. For purposes of the NIS, the definition of a *community hospital* is that used by the American Hospital Association (AHA): "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Affairs hospitals, Indian Health Service hospitals, and other Federal hospitals are excluded. Also, short-term rehabilitation hospitals were excluded beginning in 1998, and long-term acute care hospitals were excluded beginning in 2012.

The SID are the building blocks of the NIS. For each participating State,[1] the SID contain nearly the entire population of discharges from all hospitals in the State (not just community hospitals), subject to data availability and State reporting requirements. HCUP translates the SID data into a uniform format to facilitate multi-State comparisons and analyses. SID releases for data years 1990 through 2012 can be purchased for States that choose to release their data through the HCUP Central Distributor. Costs vary by State and data year.

In addition to raising users' awareness about the need for missing data methods, this report emphasizes considerations specific to the NIS and SID data. For example, the NIS is a large complex sample of discharges (Houchens et al., 2014). It is important to incorporate the sample design elements (e.g., clusters, strata, discharge weights) into the imputation models and into the analyses of the imputed data. Also, analysts are sometimes interested only in estimates for analytic subsets of the NIS (e.g., patients with diabetes). Special procedures are required for proper imputation and analyses of NIS subsets. Sample design issues do not apply to the SID.

Throughout this report, *missing data* refers to missing values for data elements, such as age, race, sex, and charges, not to entirely missing observations. *Imputation* is a procedure for replacing missing values with valid imputed values. For example, an imputation procedure

[1] As of the 2012 data year, 47 States participated in HCUP.

might replace missing values for sex with codes for male and female. Subsequent data analyses incorporate the imputed values to make statistical inferences. As discussed in this report, inferences based on the combination of nonmissing and imputed data are more likely to be valid than analyses based solely on nonmissing data.

Chapter 2 of this report describes a typology of missing data, critical to deciding whether missing data should be imputed, how they should be imputed, and how inferences should be made using the imputed data. Chapter 3 discusses missing data for selected data elements in the NIS 2012 and the Michigan SID 2012. These two databases are used to illustrate missing data methods later in this report. Chapter 4 briefly reviews methods for analyzing missing data; it emphasizes multiple imputation, which is the method that we recommend for imputation and subsequent analyses of the NIS and SID. Chapter 5 discusses imputation procedures that are available in statistical software packages. Chapter 6 contains imputation examples using the NIS and the SID. Chapter 7 presents general recommendations for handling missing data in HCUP data.

2. TYPES OF MISSING DATA

In his seminal paper on missing values, Rubin (1976) created a typology for missing values that persists to the present day. He identified three classes of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Each of these classes corresponds to a set of assumptions concerning the reasons that the data are missing—the missing data mechanism—and each class has specific implications for valid analysis and inference.

These classes are formally defined by precise mathematical probability statements concerning the missing data as they relate to both observed and unobserved data values. We omit the mathematical probability statements, which are available in numerous references, including Rubin (1976). Instead we simply describe the conditions that must be met for missing data in each class and the implications for imputing missing data.

This discussion is meant only as an introduction to these concepts. Interested readers can obtain the precise definitions and more examples from Rubin (1976), Van Buuren (2012), and Carpenter and Kenward (2013).

2.1 Missing Completely at Random

Data are missing completely at random (MCAR) if the probability of a missing value on a study unit is unrelated to both the observed and the unobserved data on that study unit. The implications are straightforward: if data are MCAR, then each unit has the same probability of a missing value, regardless of the unit's characteristics. Thus, the observed (nonmissing) values are statistically representative of the entire population of units. In this case, excluding the missing data from the analysis will result in unbiased estimates. Naturally, the estimates will suffer a loss in precision because of the reduced sample size compared with a strategy that includes the observations with missing data.

Most statistical software effectively defaults to this assumption. Although this is the simplest assumption for inference with missing data, it is often unrealistic, resulting in biased estimates. As an extreme example, suppose that we wanted to estimate in-hospital mortality for patients hospitalized with diabetes. Suppose that mortality information was missing for all Medicare patients but not missing for any other patients. Clearly, estimates based solely on the nonmissing data would underestimate mortality because Medicare patients are older, on average, and mortality risk increases with age.

2.2 Missing at Random

Data are missing at random (MAR) if the probability of a missing value (1) depends only on observed data and (2) is independent of data not observed for the unit. Another way of thinking about this is that data are MAR if they are MCAR within groups defined by the observed data. We say that a data element is MAR *dependent* on the observed groups or *dependent* on some known property. This is best explained through an example. The following example is somewhat unrealistic, but it illustrates the MAR principle.

Suppose that total charges are missing more often for higher-cost cases, which means that the probability of a missing value for total charges depends on its unobserved value. Clearly, total charge values are not MCAR. However, suppose that total charges are MCAR separately for (higher-cost) surgical and (lower-cost) nonsurgical cases. Then, if the observed data include an indicator for surgical cases, we would say that total charges are MAR, *dependent on the surgical indicator*.

Critically, the statement "total charges are MAR, dependent on the surgical indicator" is untestable with the data. This statement could be verified only if we were able to observe the missing data. However, analysts should satisfy themselves that associations used to justify the MAR assumption have a strong underlying rationale and that they are consistent with the observed data. One approach would be to test the relationships in similar datasets without missing values for the relevant data items. For instance, it might be possible to use MedPAR data to test whether estimates for Medicare patients in the NIS are consistent with those from the MedPAR data, separately for surgical and nonsurgical cases.

In summary, data are MAR if missingness depends only on observed data and not on missing data. We can use techniques described below to generate unbiased estimates and variances if the data are MAR.

2.3 Missing Not at Random

If the missing data are neither MCAR nor MAR, then they are missing not at random (MNAR). In other words, the probability of a missing value varies for reasons unknown or for reasons not encoded in the observed data. For example, a data source might set total charges to missing for all patients in swing beds. If we cannot identify discharges treated in swing beds from that particular data source, then charges are MNAR. In that case, we will need to model the missingness in order to generate unbiased estimates.

3. MISSING DATA IN THE NIS AND SID FILES

In this chapter, we tabulate missing value rates for selected variables in the NIS 2012 and Michigan SID 2012 files. This section is not intended as a comprehensive description of missing values in these files. Rather, it is intended to prompt readers to think about the types of missing values that occur in the data and about how to start researching reasons for missing values.

Documentation on general HCUP coding of missing values is available on the AHRQ Web site (http://www.hcup-us.ahrq.gov/db/coding.jsp#missing). HCUP data contain different codes for various types of missing values, including invalid, unavailable, inconsistent, and not applicable data. In addition, it is important to recognize that some variables may have valid codes that actually represent missing values. For example, the major diagnostic category (MDC) is missing whenever MDC=0, and the diagnosis related group (DRG) is missing whenever DRG=999. The missing values for MDC and DRG can result from missing or invalid principal diagnoses, wrong sex for the diagnosis or procedure, and so on. Code information for specific data elements is available on the AHRQ Web site (http://www.hcup-us.ahrq.gov/db/nation/nis/nisdde.jsp).

Some data element values are missing by design. For example, the data element PRDAY1, the number of days from admission to the first-listed procedure, is relevant only for discharge records with a procedure, and it should be considered truly missing only if the primary procedure PR1 is *not* missing.

Finally, it is impossible to identify missing values for some data elements. For example, a blank value for a secondary diagnosis (DX2-DX25) or for a procedure (PR1-PR15) presumably indicates "none," but it could also represent incomplete coding (a missing diagnosis or procedure). In that case, some of a patient's major complications, comorbidities, and procedures may have been inadvertently excluded from the discharge abstract.

3.1 Missing Data in the NIS 2012

Table 1 shows the rate of missing values in the NIS 2012 for selected data elements. These are *overall* missing value rates. The rates could be higher or lower for analytic subsets. For example, data elements could be missing at a higher or lower rate for discharge records with a specific diagnosis or a specific procedure. Also note that DIED and DISPUNIFORM have the same rate of missing values, because the data element DIED is derived from the data element DISPUNIFORM. Consequently, the best imputation strategy for DIED might be to impute it on the basis of an imputation of DISPUNIFORM. Fortunately, the overall missing value rates are far below 1 percent for all listed data elements except race (5.7 percent), total charges (2.1 percent), and median income (2.2 percent). Next, we discuss these three variables in more detail and provide insights into how to investigate missing data using HCUP documentation.

To learn more about the coding of race in the NIS 2012, we consulted detailed documentation on the AHRQ Web site (http://www.hcup-us.ahrq.gov/db/vars/race/nisnote.jsp#general). According to the State-specific notes, race was suppressed in California for some discharges with sensitive conditions (e.g., HIV and AIDS). Therefore, in California, race values are not

missing completely at random. Also, Louisiana does not collect Hispanic ethnicity information, and in Utah a large hospital system stopped collecting Hispanic ethnicity information, meaning that most Hispanic patients were presumably coded as either White or Black. Consequently, race is coded differently (but not missing) for all discharges in Louisiana and for some discharges in Utah. Finally, race was not reported for some entire states: Minnesota, North Dakota, and West Virginia.

The patient's State was dropped as a data element from the NIS 2012 file to enhance patient confidentiality. Therefore, California, Louisiana, Utah, Minnesota, North Dakota, and West Virginia discharges cannot be singled out in the NIS 2012 for special treatment of the race variable. The California missing values should not be an issue for nonsensitive diagnoses and conditions. For example, discharge records with a principal diagnosis of diabetes would be largely unaffected by California's suppression of race for cases with sensitive conditions. Although the NIS 2012 does not identify State, it does identify the census division, which might prove helpful for imputation, along with general population race information at the State and census division level available from the U.S. Census Bureau.

By again consulting the AHRQ Web site documentation, we found that total charges in the NIS 2012 are assigned specific missing value codes if the reported value is equal to zero, less than $25, or more than $5 million. Some California hospitals were exempt from reporting total charges (including all Kaiser and Shriner hospitals); charges for discharges from these hospitals were set to missing. Reporting total charges was optional for Kansas hospitals. Total charges were not reported for Maryland hospitals. Ohio excluded total charges that were zero, under $100, or more than $1 million. In Oregon, some hospitals did not report total charges for charity care and Kaiser Hospitals were exempt from reporting total charges. In Wisconsin, hospitals were not required to report total charges for stays longer than 100 days.

Therefore, total charges are not missing completely at random from the overall NIS 2012 data. Missing values are often assigned to total charges if the *original* data values were very low or very high. Except for charges in Wisconsin, one could take the position that these extreme original values were mostly coding errors unrelated to the true values. Kaiser hospitals in California and Oregon tend to serve members of Kaiser Health plans (health maintenance organizations). It is possible that total charges for discharges from these hospitals would differ systematically if they were compared with discharges from non-Kaiser hospitals; however, it is impossible to assess these differences from the observed data. On the assumption that systematic differences in total charges would be reflected by similar differences in length of stay (LOS), one imputation strategy would be to predict total charges from LOS, which is missing for only 0.03 percent of all discharges.[2]

[2] To impute charges from LOS, one could first impute LOS for the 0.03 percent of discharges with missing values.

Table 1. Percentage of missing values for selected data elements, NIS 2012

Data Element	Label	% Missing (N=7,296,968)
AGE	Age in years at admission	0.05
AMONTH	Admission month	0.13
AWEEKEND	Admission day is a weekend	0.03
DIED	Died during hospitalization	0.02
DISPUNIFORM	Disposition of patient (uniform)	0.02
DQTR	Discharge quarter	0.10
DX1	Diagnosis 1	0.07
ELECTIVE	Elective versus non-elective admission	0.34
FEMALE	Indicator of sex	0.01
HCUP_ED	HCUP Emergency Department service indicator	0.00
HOSPBRTH	Indicator of birth in this hospital	0.00
LOS	Length of stay (cleaned)	0.03
PAY1	Primary expected payer (uniform)	0.25
PL_NCHS2006	Patient location: NCHS Urban-Rural Code (V2006)	0.41
RACE	Race (uniform)	5.73
TOTCHG	Total charges (cleaned)	2.08
TRAN_IN	Transfer in indicator	0.54
TRAN_OUT	Transfer out indicator	0.02
ZIPINC_QRTL	Median household income national quartile for patient ZIP Code	2.23

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), National Inpatient Sample (NIS), 2012

The median household income quartile (ZIPINC_QRTL) was set to missing whenever the patient's ZIP Code was missing or when it could not be matched to a ZIP Code in the data source that provides median household income. To protect patient confidentiality, median income was also set to missing for all discharges in ZIP Codes with populations below a minimum threshold and for all discharges in a ZIP Code that solely represented an entire income quartile for its State. Clearly, the income quartile is not missing at random because it is often missing for all discharges within specific ZIP Codes. Ideally, we would impute the income quartile from ZIP Code level data so that a single income quartile would be assigned to all discharges residing within a given ZIP Code. However, to protect patient confidentiality AHRQ cannot provide patient ZIP Codes in the NIS. Although it will not be possible to ensure that imputed values are the same value for each ZIP Code, perhaps relationships can be exploited between the income variable and other NIS variables such as race and primary payer (PAY1).

3.2 Missing Data in the Michigan SID 2012

Table 2 shows the rate of missing values in the 2012 Michigan SID for selected data elements. Again, these are *overall* missing value rates, and the rates could be higher or lower for analytic subsets. A few things stand out about the entries in Table 2 compared with those in Table 1.

Table 2. Percentage of missing values for selected data elements, Michigan SID 2012

Data element	Label	% Missing (N=1,249,805
AGE	Age in years at admission	0.01
AMONTH	Admission month	0.00
ATYPE	Admission type	0.36
AWEEKEND	Admission day is a weekend	0.00
DIED	Died during hospitalization	0.02
DISPUB04	Disposition of patient (UB-04 standard coding)	0.02
DISPUNIFORM	Disposition of patient (uniform)	0.02
DQTR	Discharge quarter	0.00
DaysCCU	Days in coronary care unit (as received from source)	0.00
DaysICU	Days in medical/surgical intensive care unit (as received from source)	0.00
DX1	Diagnosis 1	0.01
DXPOA1	Diagnosis 1, present on admission indicator	0.37
DX_ADMITTING	Admitting diagnosis code	9.00
FEMALE	Indicator of sex	0.01
LOS	Length of stay (cleaned)	0.01
LOS_X	Length of stay (as received from source)	0.00
MDNUM1_R	Physician 1 number (reidentified)	4.52
MEDINCSTQ	Median household income state quartile for patient ZIP Code	0.64
PAY1	Primary expected payer (uniform)	0.07
PL_CBSA	Patient location: Core Based Statistical Area (CBSA)	0.11
PL_MSA1993	Patient location: Metropolitan Statistical Area (MSA), 1993	0.11
PL_NCHS2006	Patient location: NCHS urban-rural code (V2006)	0.11
PL_RUCA10_200	Patient location: Rural-urban commuting area (RUCA) Codes, 10 levels	2.02
PL_RUCA2005	Patient location: Rural-urban commuting area (RUCA) Codes	2.02
PL_RUCA4_2005	Patient location: Rural-urban commuting area (RUCA) Codes, 4 levels	2.02
PL_RUCC2003	Patient location: Rural-Urban Continuum Codes (RUCCs), 2003	0.11
PL_UIC2003	Patient location: Urban Influence Codes, 2003	0.11
PL_UR_CAT4	Patient Location: Urban-rural, 4 categories	0.11
PROCTYPE	Procedure type indicator	0.00
PSTCO2	Patient State/county FIPS code, possibly derived from ZIP Code	0.10
RACE	Race (uniform)	17.15
TOTCHG	Total charges (cleaned)	19.79
TOTCHG_X	Total charges (as received from source)	19.77
TRAN_IN	Transfer in indicator	0.84
TRAN_OUT	Transfer out indicator	0.02
ZIPINC_QRTL	Median household income national quartile for patient ZIP Code	0.64
ZIP_S	Patient ZIP Code (synthetic)	0.03

Abbreviations: NCHS, National Center for Health Statistics; FIPS, Federal Information Processing Series.

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), State Inpatient Databases (SID) for Michigan, 2011

First, the SID contain some data elements that are not available in the NIS. For example, an encrypted physician code is available in the Michigan SID and, for some data elements, the SID contain both the uniformly coded (NIS) value and the value from the source data. Second, several data elements have identical missing value rates. Data elements with identical missing value rates are usually derived from a common data element. For example, most of the Patient Location variables (data element names beginning with "PL") are derived from patient ZIP Code. Therefore, the place variables are missing whenever the patient ZIP Code is missing. Care should be taken to ensure that imputed values for related data elements such as these are consistent with one another. For example, for a single imputation, it would be unfortunate if one imputed variable implied a rural location and another simultaneously imputed variable implied a metropolitan location.[3]

Fortunately, the overall missing value rates for the Michigan SID are far below 1 percent for most of the listed data elements. Exceptions include the admitting diagnosis code (9.0 percent), the physician number (4.5 percent), and the rural-urban commuting area (RUCA) codes (2.0 percent). The missing value rates for race in the 2012 Michigan SID are much higher than in the NIS 2012 (17.2 percent vs. 5.7 percent), and the missing value rates for total charges are also much higher in the 2012 Michigan SID than in the NIS 2012 (19.8 percent vs. 2.1 percent).[4]

There are more than 10,000 ICD-9-CM diagnosis codes, so imputing the admitting diagnosis is problematic. Fortunately, very few studies use the admitting diagnosis. That said, for those few studies it might be possible to reasonably impute the admitting diagnosis by exploiting its relationship with the principal diagnosis and other data elements.

The (encrypted) physician number is missing on 4.5 percent of records, and there is no way to reasonably impute missing values using the available data. Analysts who rely on the physician number should compare statistics from records with and without missing physician numbers to determine whether the values appear to be missing more or less often for some types of discharges.

RUCA codes are missing for only 2 percent of the records. Analysts who want to impute RUCA codes should consider this in a vein similar to that discussed in the previous section on the imputation of income quartiles.

The information in the State-specific notes on missing race values for the Michigan SID 2012 (http://www.hcup-us.ahrq.gov/db/vars/siddistnote.jsp?var=race) and on missing total charges (http://www.hcup-us.ahrq.gov/db/vars/siddistnote.jsp?var=totchg) do not give us any clues about the reasons that these data elements are sometimes missing. Therefore, to develop an imputation strategy for these data elements, analysts should learn what they can about missing value patterns and relationships directly from the SID data. To the extent that the race distribution of 2012 Michigan discharges mirrors the race distribution of the 2012 Michigan

[3] This consistency should hold *within* each single imputation, but not necessarily *across* multiple imputations, if multiple imputation is used.

[4] Because the NIS is a sample of discharges from the collection of HCUP SID files, the missing value rate for total charges in the NIS is partly a function of the missing value rate for total discharges in the Michigan SID. An analyst may find such information useful when thinking about missing charges in the NIS.

population, population data from the U.S. Census Bureau might be helpful for imputing missing values for race.

4. MISSING DATA METHODS

This chapter contains a high-level overview of commonly applied missing data methods. Analysts who are interested in the theoretical details and statistical properties of these methods should consult the references given in this chapter and at the end of this report. We start with brief descriptions of some commonly used methods that should be avoided. We then touch briefly on maximum likelihood methods and conclude this chapter with multiple imputation methods, which are generally recommended.

4.1 Case Deletion Method

The default behavior for most statistical software is to delete cases with missing values. Although this is the most expedient method for handling missing values, it yields unbiased estimates only if the data are MCAR. Otherwise, this method should be avoided.

4.2 Single Imputation Methods

Single imputation replaces each missing value once, after which the analysis proceeds as though the data were complete. Historically, many methods have been used for single imputation. Most of these methods have severe limitations, which is one reason why AHRQ does not provide single imputed values for the NIS and SID data. Among the most popular methods for single imputation are mean imputation, missing value indicators, and last observation carried forward. Later, we discuss a method called predictive mean matching, which the Medical Expenditures Survey uses for single imputation.

4.2.1 Mean Imputation

Mean imputation is performed by setting missing values equal to the mean of the nonmissing values. Regardless of whether the imputed value is based on an overall mean, a mean of subsets (e.g., separate means for males and females), or a (mean) regression estimate, this method suffers two drawbacks. First, the mean of the missing values (if they could be observed) might be different from the mean of the nonmissing values. Second, the sample variance will be artificially deflated because of the low (or nonexistent) variation among imputed values. In turn, significance levels will be artificially inflated, potentially causing wrong inferences. Consequently, this imputation method is not recommended.

4.2.2 Missing Value Indicators

For regressions, missing predictors have been handled by (1) assigning a value of zero to the missing predictor and (2) adding a missing value indicator (equal to 1 if the predictor is missing and equal to 0 otherwise) to the set of predictors. Thus, the regression "adjusts" predictions for cases with missing values. Although this method has the advantage that it retains all of the data in the analysis, it is not recommended because it can result in biased estimates even if the data are MCAR (van Buuren, 2012).

4.2.3 Hot-Deck and Cold-Deck Imputation

Hot-deck imputation replaces missing values with nonmissing values taken from a randomly selected, closely matched observation in the same data set as the observation containing the missing value. Cold-deck imputation replaces missing values from observations matched in a different data set. Andridge and Little (2010) discuss hot-deck imputation methods that can be used successfully for single imputation.

4.2.4 Last Observation Carried Forward

One form of hot-deck imputation sometimes used for longitudinal and time series data, replaces missing values with the previous nonmissing value in the time series. This method might make sense for data elements that do not vary over time, such as sex. However, more generally this method can result in biased estimates even if the data are MCAR. Although it is an acceptable method for clinical trials under certain circumstances (National Research Council, 2010, p. 77), it is not normally recommended for observational data such as those available from HCUP.

4.3 Weighting Methods for Finite Population Statistics

For finite population estimates based on data from sample surveys, missing cases (survey nonrespondents) or cases with missing data are often reweighted according to the sample design weights and perhaps adjusted for nonresponse bias. Most inferences treat the NIS and SID as samples from infinite rather than finite populations (Houchens, 2010). Therefore, readers interested in this approach are referred to sampling texts such as Cochran (1977) and Foreman (1991).

4.4 Likelihood-Based Methods

Likelihood-based methods stipulate a model for the *observed* data, resulting in an *observed* likelihood. The parameters are often estimated by maximizing the *observed* likelihood using the expectation-maximization (EM) algorithm or by Bayesian simulation of the posterior distribution. We discuss both methods briefly in this section. These methods can be useful, but we do not discuss them further because custom-written or specialized software is often required.

4.4.1 Expectation-Maximization Algorithm

The EM algorithm is an iterative method for maximizing the observed likelihood. We start with crude estimates for the parameters, perhaps based on the nonmissing data, and then iterate between two steps: (1) (E-step) computing expected values for the missing data given the current parameter estimates and (2) (M-step) estimating new parameter values by maximizing the observed likelihood after substituting the values estimated in step 1 into the likelihood equation. Details and examples of the EM method can be found in the books by Little and Rubin (2002), Kim and Shao (2014), and Gelman et al. (2014), among other references.

4.4.2 Bayesian Estimation

Bayesian methods estimate values for missing data just as they estimate any other unknown model parameter. A joint model is set up for the observed data, the unobserved data, and the parameters. Parameter values are then estimated from the simulated joint posterior distribution *conditional on the observed data*. One approach is to set the model up in three parts: a prior

distribution for the parameters, a joint model for all of the data (missing and observed), and a model for the missingness process. Readers who are interested in the Bayesian approach can consult many books on Bayesian data analysis, including Congdon (2006) and Gelman et al. (2014).

4.5 Multiple Imputation Methods

Multiple imputation can be considered a flexible extension of likelihood-based methods with the added benefit that it is often easier than likelihood-based methods to obtain good estimates of standard errors for a wide range of parameters, which is critically important for statistical inferences. There are several good books on multiple imputation. The two most recent books are by Carpenter and Kenward (2013) and Van Buuren (2012). Here we give a brief overview of multiple imputation to introduce the key concepts.

Multiple imputation can be used for data MAR, and it can help bring data MNAR closer to MAR (Gelman and Hill, 2007). According to Van Buuren (2012, p. 27), "Nowadays multiple imputation is almost universally accepted, and in fact acts as the benchmark against which newer methods are being compared." *It is for these reasons and because multiple imputation has been made available in all of the major statistical packages that we recommend it as the method of choice for most NIS and SID studies*.

Multiple imputation has its roots in some work that its founder, Donald B. Rubin, performed to address missing income data in the Current Population Survey in the 1970s. Rubin recognized, among other things, that imputing one value (single imputation) does not allow for the uncertainty associated with that one value when one is calculating standard errors of the statistics generated from the imputed data.[5]

His solution was to produce multiple versions of a dataset, each with a single imputed value for the missing data elements. The imputed values were generated in a way that varied across the multiple versions to reflect the uncertainty associated with each of the imputations. In other words, he generated one imputation for each missing value to create one "complete" version of the data, then he generated a second imputation for each missing value to create another "complete" version, and so on. Later, he developed "Rubin's rules" for combining the complete data estimates across the multiple versions, and he stated the conditions under which valid inferences could be made from estimates based on multiple imputations (Rubin, 1987).

Generally, multiple imputation involves three steps:

1. Define the imputation models and use them to generate M versions of "complete" data. Each of the M complete data versions has every missing value replaced by an imputed value that is plausible for that data element. The imputed values for each data element vary over the M versions, reflecting the uncertainty of the imputed values. There are several options for the imputation models, and we describe some of the most common imputation models in the subsections that follow.

[5] However, appropriate variance formulas have been developed since that time, making single imputation a viable option for some analyses (Rao and Shao, 1999; Andridge and Little, 2010).

2. For each of the M versions of complete data, estimate the parameters of interest using whatever statistical methods would have been used if the original data had been complete (no missing data values). For example, the parameters of interest might be means, variances, correlations, regression coefficients, and so on. The estimates will differ across the M versions solely because the imputed values differ across the M versions. This yields M estimates for each parameter of interest.
3. Pool the M imputation estimates to calculate a single estimate for each parameter of interest and calculate its variance using Rubin's rules. The variance incorporates both the within-imputation variance and the between-imputations variance, reflecting the uncertainty associated with the imputations. Under suitable conditions, the pooled estimates are unbiased and yield better inferences than do estimates based only on the nonmissing data values.

One item that must be decided is the number M of imputations (complete data versions) to generate. Recommendations tend to range between 5 and 100, sometimes depending on the parameters of interest and the degree of missing data (Carpenter and Kenward, 2013). Given that the only penalty for generating more imputations is increased usage of computer time and disk space, Berglund and Heeringa (2014) recommend a small number of imputations for the exploratory phase (M=5–10) and recommend a larger number for the final analysis (M=30 to 100). Another approach is to increment M until the results "stabilize" to ensure a sufficient level of statistical efficiency.

4.5.1 Common Imputation Models

An even more critical element that must be chosen is the imputation model. This choice will depend on the types of data elements in the analysis—whether they are binary, multinomial, continuous, a count, or a mixture of types—and on the joint and marginal distributions of the data elements.

These models often involve regression inputs (predictor variables). The conventional advice is to include as many predictors as possible. It is impossible to prove that data are MAR, so we should "hedge our bets" by including all variables and combinations of variables (e.g., interactions) that might affect a variable's probability of missingness to make the MAR assumption more plausible (Gelman and Hill, 2007). It is important to remember that the goal is accurate prediction, not causal inference. Thus, it is acceptable to use any predictors that are available.

It is important for the imputer to ensure that the imputation model and the data model share a common analytic goal (Meng, 1994; Meng, 2000). This can be especially important if the imputation and analysis are done separately by different individuals or by different organizations.

In the following subsections, we describe several common imputation models and indicate the types of variables for which each is relevant.

4.5.1.1 Regression Imputation

Regression imputation is useful for imputing continuous variables, such as total hospital charges. If Y represents the continuous variable with missing values to be imputed and **X** represents a vector of predictor variables, then a linear regression model is fit (usually) by the method of ordinary least squares (OLS). All of the usual assumptions concerning OLS regression apply.

The idea is to estimate the regression coefficients and the error variance for a regression model. The errors are assumed to be normally distributed so that the regression coefficients and the error variance have known statistical distributions. Imputations are generated on the basis of predictions generated by random draws from the statistical distributions for the coefficients and the sampling variance.

One common issue with OLS regression is that "impossible" predictions are possible. For example, a regression for total hospital charges might produce imputations with negative values for some observations. This might not be of great concern if the objective is simply to estimate average charges over a large sample. The mean might still be unbiased. However, in the event that negative imputations are problematic, the analyst might choose to fit a log-linear model to ensure positive predictions. Other solutions replace negative values with "minimum" positive values. However, that procedure will most likely bias the estimates. Von Hippel (2013) offers a detailed discussion of the effects of these procedures and of transformations to correct skewness.

4.5.1.2 Predictive Mean Matching Imputation

Predictive mean matching imputation is suitable for imputing values for most types of variables. It is similar to regression imputation, but the regression method is specific to the variable type. For example, we would employ logistic regression for binary outcomes. The difference from regression imputation is that the imputed values for a variable are drawn only from the observed (nonmissing) values for that data element. In particular, a value is randomly drawn from observed values that are within a "neighborhood" of the regression prediction. The neighborhood is usually defined as the K nonmissing values that are closest to the regression-predicted value. This process ensures that all imputations will be within the range of the observed values. For example, if K=5 and the regression predicted a negative value for hospital charges, then the imputation would be randomly drawn for the five lowest observed values for hospital charges.

Predictive mean matching is successfully used as a *single* imputation strategy for several items with missing values in the Medical Expenditures Panel Survey.[6] However, this strategy requires special variance calculations that are not standard in the major statistical packages (Rao and Shao, 1999; Robins and Wang, 2000; Andridge and Little, 2010).

---

[6] See for example, Agency for Healthcare Research and Quality, Center for Financing Access, and Cost Trends. MEPS HC-144F:2011 Outpatient Department Visits. Rockville, MD: Agency for Healthcare Research and Quality; July 2013. http://meps.ahrq.gov/mepsweb/data_stats/download_data/pufs/h144f/h144fdoc.pdf. Accessed December 4, 2014.

4.5.1.3 Propensity Score Imputation

Propensity score imputation uses logistic regression to estimate the probability that a data element is missing. The idea is then to draw imputed values from the nonmissing observations with a propensity score that is similar to the observation with the missing value. Berglund and Heeringa (2013) advise strongly against this method if the objective is a multivariate analysis because associations among variables are not preserved. That said, it could be useful for univariate analysis.

4.5.1.4 Logistic Regression Imputation

Logistic regression imputation is suitable for imputing missing values for binary variables (e.g., male/female, 0/1, or yes/no responses). Some forms of logistic regressions can also handle ordinal and nominal (multinomial) responses.

As with the regression method, for a binary response a logistic regression model is fit to predict a "yes" response on the basis of available predictors. This results in estimates of the logistic regression coefficients, which are assumed to be normally distributed with an estimated covariance matrix. For each observation with a missing response, the probability of a "yes" response is estimated from the logistic regression on the basis of a random draw of the regression coefficients from this distribution. Assume, for example, that the estimated probability of a "yes" response is equal to 0.15. Then a "yes" response is imputed for the missing value with 15 percent probability and a "no" response is imputed with 85 percent probability.

4.5.1.5 Discriminant Function Imputation

Discriminant function imputation is suitable for imputing missing values for multinomial variables (unordered discrete response categories such as patient race).

This method is similar to the other regression procedures, except that it predicts group probabilities (such as the probability that race is White, Black, Hispanic, and so on) on the basis of "predictor" variables. First the group probabilities for a given observation are simulated from the fitted model. Then the response is imputed with the simulated probabilities. For example, say we had three responses (red, white, and blue) with simulated probabilities (0.5, 0.2, 0.3) for a given observation. Then the missing value would be imputed as red, white, or blue with probabilities equal to 50 percent, 20 percent, and 30 percent, respectively.

The main difficulty with this method is that for each response category, the predictors are assumed to be distributed as multivariate normal with a common covariance matrix, which is problematic for noncontinuous predictors.

4.5.2 Multivariate Imputation

We perform *univariate imputation* when we impute missing values for only a single data element. We perform *multivariate imputation* when we impute missing values for multiple data elements. Do not confuse *single imputation* with *univariate imputation* or confuse *multiple imputation* with *multivariate imputation*. Single imputation creates a single complete dataset, and univariate imputation imputes missing values for a single variable. Multiple imputation

creates multiple complete datasets, and multivariate imputation imputes missing values for multiple variables.

All of the common imputation methods covered earlier can be used both for univariate imputation and for some kinds of multivariate imputation. In the following subsections, we discuss three multivariate imputation methods.

4.5.2.1 Markov Chain Monte Carlo Imputation

As implemented in most major statistical packages, the Markov chain Monte Carlo (MCMC) method of imputation *is suitable for multivariate normal data with or without a monotone missing data pattern*. This is a Bayesian method that uses an iterative MCMC algorithm to simulate draws from an estimate of the posterior distribution. Although some MCMC implementations allow for noncontinuous data by setting limits on imputed values and by rounding them to integer values, setting limits and rounding are not recommended for imputing large quantities of missing values on noncontinuous variables (Allison, 2005).

The MCMC algorithm iterates until the estimate of the posterior distribution stabilizes. The algorithm begins by substituting starting values for the parameters, either user supplied or algorithm supplied, and then updates parameter estimates at each iteration. The user can request multiple chains, each with a different set of starting values, to assess the influence of the starting values. For each chain, the user specifies the initial number of "burn-in" iterations, for which the estimates are discarded because the distribution has not yet stabilized. Typically, the burn-in consists of the first 1,000 to 20,000 iterations.

Convergence is assessed by whether the sequence of MCMC estimates varies around a fixed level (the mean) with a fairly constant variance over the sequence for each chain. Typically, users require between 10,000 and 100,000 iterations after the burn-in period. It also may be important to assess the autocorrelation in the sequence of estimates. If estimates tend to be correlated over successive iterations, then either the sequence of estimates can be "thinned" by selecting every kth estimate in the series (to make the estimates uncorrelated) or a larger number of MCMC estimates can be generated to overcome the effect of autocorrelation on the precision of the estimates.

4.5.2.2 Full Conditional Specification Imputation

The full conditional specification (FCS) approach is good for imputing missing data for large mixed sets of continuous, nominal, ordinal, count, and semicontinous variables (Berglund and Heeringa, 2014; Carpenter and Kenward, 2013; Van Buuren, 2012). In other words, it can be used for imputing all types of data. Other names in the literature for this approach are *imputation using chained equations* (ICE) and the *sequential regression algorithm*.

This algorithm moves one by one through the variables to be imputed, with each variable being imputed by an appropriate method (e.g., one of the methods described earlier). For example, linear regression might be used to impute continuous variables and logistic regression might be used to impute binary variables. After all of the variables have been imputed once, the algorithm cycles through another round of imputations, during which the previous imputations

are treated as real, nonmissing data. The algorithm iterates until the estimates converge, and the final imputations are generated from the converged distribution.

4.5.2.3 Monotone Imputation

More expedient imputation algorithms can be applied if the missing data pattern is *monotone*. The missing data pattern is monotone if the list of data elements with missing values can be reordered so that (1) if the jth element in the ith record is not missing, then all previous data elements in the list are also not missing and (2) if the jth element in the ith record is missing, then all subsequent data elements in the list are also missing. The diagram below illustrates a monotone missing data pattern.

Observation	Data Elements			
	Y1	Y2	Y3	Y4
1	X	.	.	.
2	X	X	.	.
3	X	X	X	.
4	X	X	X	X
5	X	X	X	X

For each of five observations, the diagram shows whether data are missing for each of four data elements Y1, Y2, Y3, and Y4 (e.g., age, sex, race, and LOS). An "X" indicates that the data element is not missing, and a "." indicates that the data element is missing. The observations and data elements have been ordered in such a way that a missing value in one cell is always followed by missing values in the cells to the right of it. Conversely, a nonmissing value in a cell is always preceded by nonmissing values in the cells to the left of it. Note that Y1 happens to be nonmissing for all of the observations.

The imputation algorithm can take advantage of this pattern by first imputing Y2 based on Y1, then imputing Y3 based on Y1 and Y2, and so on. Any of the imputation methods appropriate to the missing data can be applied to the sequence of missing data elements.

Missing data patterns that are not monotone are called *arbitrary*. Arbitrary missing data patterns can sometimes be converted to monotone patterns by first imputing a small amount of the missing data to produce a monotone pattern.

5. MISSING DATA SOFTWARE

All of the major statistical packages now implement some forms of multiple imputation. The software list seems to grow every year. Van Buuren maintains a current list of multiple imputation software on his Web site: http://www.stefvanbuuren.nl/mi/Software.html.

Although specific MCMC methods are implemented in most of the major statistical packages, most of the more general Bayesian approaches to missing data are implemented in Bayesian software such as WinBUGS (http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/), OpenBUGS (http://www.openbugs.net/w/FrontPage), and Stan (http://mc-stan.org/), which are all freely available for download. Further, utilities have been developed for these

packages so that they can all be executed from the R language. Macros have also been developed to execute WinBUGS and OpenBUGS from SAS.

The examples in the next chapter all use SAS to perform multiple imputations. We briefly describe multiple imputation using three of the most popular statistical packages: SAS, Stata, and R. All of the multiple imputation methods discussed earlier are available in each of these three statistical packages.

5.1 SAS

SAS implements multiple imputation in three steps. First, PROC MI is used to generate M copies of imputed data into a single file, where the number of each imputation 1, 2… M, is stored in a variable called _imputation_. Second, the user obtains separate estimates for each value of _imputation_, usually through the use of a "BY _IMPUTATION_" statement with the procedure chosen for analysis (e.g., PROC MEANS, PROC REG). Third, PROC MIANALYZE is used to combine the M estimates into a single estimate for each parameter along with an estimate of its variance using Rubin's rules.

SAS users should consult the SAS Stat Users Guide for PROC MI and PROC MIANALYZE, which are available at http://support.sas.com/documentation/onlinedoc/stat/ex_code/121/. Also, the book *Multiple Imputation of Missing Data Using SAS* by Berglund and Heeringa (2014) provides an excellent introduction to multiple imputation that offers advice and real-world examples for performing multiple imputation using SAS. The capabilities for imputation in SAS are further broadened by the availability of the user-written SAS macro *IVEware* (Raghunathan et al., 2002), which can be downloaded for free (http://www.isr.umich.edu/src/smp/ive/). Finally, for SAS users who are familiar with the R language, all of the imputation packages in R can be essentially implemented in SAS through a user-written SAS macro to execute the R language within SAS (Wei, 2012).

5.2 Stata

Stata offers functions to implement all of the multiple imputation methods discussed in the previous section (StataCorp, 2011).

First, data are imputed using the MI IMPUTE command. Univariate imputation models include linear regression and predictive mean matching for continuous variables, as well as truncated regression and interval regression for continuous variables that have a restricted range or are censored. Imputation models also include logistic regression for binary variables, ordered logistic regression for ordinal variables, and multinomial logistic regression for nominal variables. Finally, Poisson regression and negative binomial regression can be used for count variables. Multivariate imputation models include fully conditional specification (FCS), monotone methods, and multivariate normal regression.

Second, the analysis and pooling steps are performed by the MI ESTIMATE command. It runs a user-specified estimation procedure on the data produced from MI IMPUTE and then pools the estimates using Rubin's rules. It uses survey estimation commands to produce estimates from survey data using sample design factors.

Stata also has routines for displaying missing value patterns. Details are available on the Stata Web site: http://www.stata.com/features/multiple-imputation/ and in their *Multiple-Imputation Reference Manual*: http://www.stata.com/bookstore/multiple-imputation-reference-manual/index.html, which can be downloaded free of charge. A worked example of multiple imputation is included in *A Gentle Introduction to Stata* by Aycock (2014).

5.3 R

R is a popular statistics language, especially among academics, because it is free and it provides comprehensive data management tools along with up-to-date user-contributed statistical routines (R Development Core Team, 2011). R offers all of the imputation methods that SAS and Stata offer. R may be downloaded from the R Web site: http://www.r-project.org/.

R offers many "packages" for imputation, including single imputation (e.g., packages *impute* and *imputation*) and multiple imputation (e.g., packages *MI, mice, VIM, Amelia,* and *Zelig*). R is the first place to look for implementations of the newest imputation algorithms. The article titled "Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box" by Su et al. (2011) is an excellent guide to imputation in general and to the use of the MI package in particular. The book *Flexible Imputation of Missing Data* by Stef van Buuren (2012) is another good reference for multiple imputation with R because it uses the R package mice (developed by Van Buuren) throughout the book. There are many other books related to R listed on the R Web site.

6. EXAMPLES

The examples in this chapter are intended to illustrate the process by which data can be imputed in the SID and the NIS, but these analyses are not definitive. For instance, the imputation models could certainly be expanded to incorporate more variables, use different variable transformations, and so on. Moreover, other imputation methods could be applied. In fact, it is often a good idea to test the sensitivity of results by comparing the estimates derived from different approaches. Nevertheless, these examples should provide users with a good starting point for their own imputation problems. We used SAS Version 9.4 to perform all analyses, and we include the SAS code in the Appendices for users who want to perform the same or similar analyses using SAS. These same procedures can be carried out using most major statistical packages.

Rather than analyze the entire NIS and SID, we analyzed the subset of discharges with a principal diagnosis of complicated diabetes. We did this for three reasons. First, the diabetes subset is a fraction of the entire file, which reduced the computer time required for imputation. Second, outcomes such as mortality, lengths of stay (LOSs), and total charges are more homogeneous within subgroups than they are across the entire file, making the imputation for those outcomes less complicated. Third, and most important, users have often been interested in subgroup analysis using either the NIS or the SID. For example, analysts are frequently interested in making inferences about groups of patients defined by specific diagnoses or procedures. For the NIS, analysis is slightly more complex because of the sample design, and the NIS examples illustrate that extra bit of complexity.

6.1 SID Examples

In this section, we used the 2014 Michigan SID data. In section 3.2, we discussed and provided an overview of missing value rates for selected data elements. For both examples, we imputed missing values for sex, race, primary payer, and total charges. The first example represents a descriptive analysis, and the second example represents a more sophisticated regression analysis. Appendix A contains the SAS code for these examples.

6.1.1 SID Average Charges for Complicated Diabetes

Our goal was to estimate the mean and a 95 percent confidence interval for total charges for discharges with complicated diabetes. We used age, sex, race, LOS, number of chronic conditions, coronary care unit (CCU) days, intensive care unit (ICU) days, and primary payer to impute the 23 percent of discharges with missing values for total charges. Sex, race and primary payer were also missing for some observations. Therefore, we also needed to impute missing values for those data elements.

Table 3 shows the missing data pattern in the 2012 Michigan SID for the 18,590 discharges with complicated diabetes identified using AHRQ's Clinical Classification Software (DXCCS1=50). The missing values produced eight distinct missing value groups, numbered in the first column. There are nine columns representing nine data elements corresponding to age, LOS, number of chronic conditions (NCHRONIC), CCU days, ICU days, sex (FEMALE), primary payer (PAY1), race, and total charges (TOTCHG). An "X" means that the data element was not missing in the corresponding group. A "." means that the data element was missing. The final two columns contain the frequency (N) and percentage of cases for each missing value group. For the noncategorical data elements, the mean of the nonmissing values is shown in Table 4.

Table 3. Missing data patterns for complicated diabetes, Michigan SID 2012

Group	AGE	LOS	NCHRONIC	Days CCU	Days ICU	FEMALE	PAY1	RACE	TOTCHG	N	%
1	X	X	X	X	X	X	X	X	X	12,229	65.78
2	X	X	X	X	X	X	X	X	.	3,926	21.12
3	X	X	X	X	X	X	X	.	X	2,058	11.07
4	X	X	X	X	X	X	X	.	.	346	1.86
5	X	X	X	X	X	X	.	X	X	9	0.05
6	X	X	X	X	X	X	.	X	.	1	0.01
7	X	X	X	X	X	X	.	.	X	20	0.11
8	X	X	X	X	X	.	X	X	X	1	0.01
Total missing (%)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.01)	30 (0.16)	2,424 (13.04)	4,273 (22.99)	6,361 (34.22)	

Abbreviations: LOS, length of stay; CCU, coronary care unit; ICU, intensive care unit.

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), State Inpatient Databases (SID) for Michigan, 2012

Table 4. Missing data patterns for complicated diabetes, group means, Michigan SID 2012

Group	AGE	LOS	NCHRONIC	Days CCU	Days ICU	FEMALE	PAY1	RACE	TOTCHG ($)	N
1	50.7	4.6	6.9	0.08	0.29	0.47	-----	-----	23,224	12,229
2	50.7	4.2	6.7	0.03	0.23	0.49	-----	-----	.	3,926
3	49.1	5.5	6.80	0.03	0.19	0.45	-----	.	27,230	2,058
4	51.4	4.2	7.6	0.03	0.05	0.44	-----	.	.	346
5	49.4	4.4	3. 7	0	0.22	0.67	.	-----	13,822	9
6	16.0	1.0	1.0	0	0	1.00	.	-----	.	1
7	33.6	4.0	6.2	0	0	0.40	.	.	16,545	20
8	51.0	5.0	8.0	0	0	.	-----	-----	13,245	1

Abbreviations: LOS, length of stay; CCU, coronary care unit; ICU, intensive care unit. Dashes indicate nonmissing nominal variables for which averages are meaningless.

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), State Inpatient Databases (SID) for Michigan, 2012

Group 1 is the group of discharges with complete data on all nine data elements, representing 12,229 (65.78 percent) of the 18,590 discharges. Thus, about two-thirds of the diabetes patients were not missing any of the nine data elements, and about one-third were missing at least one of the nine data elements. Age, LOS, chronic conditions, CCU days, and ICU days were not missing for any of the complicated diabetes discharges. Female was missing for only one discharge (group 8), and primary payer was missing for only 30 discharges (groups 5, 6, and 7). Although we could have discarded the 31 observations missing these two data elements, there was no good argument for excluding them because we could impute them easily alongside race and total charges. If they were the only two data elements with missing values, then we could consider omitting them from the analyses if it could be argued that they were MCAR or that their effects would be trivial.

The missing value rates were fairly high for race (13 percent missing overall) and total charges (23 percent missing overall). Note also, for example, in Table 4 that the observed mean LOS varied among the first three groups. The observed mean total charge for group 3 was higher than the observed mean charge for group 1, mirroring the differences in the observed mean LOS between the two groups. This hints at potentially important differences in total charges between cases with and without missing race, and it indicates that LOS might be especially helpful for imputing missing values for total charges.

The missing value pattern is not monotone (Table 3). For example, groups 2 and 3 indicate that some observations have missing charges but not missing race, and other observations have missing race but not missing charges. The missing values are for a mixture of binary, categorical, and continuous variables. Consequently, the FCS method is an appropriate multivariate imputation technique for this problem (Berglund and Heeringa, 2014).

We found that the effect of age often had inflections at age 18 and age 65. Regression splines are designed to fit flexible, continuous nonlinear functions for regression predictors. Therefore, we created a cubic spline function for age with knots at 18 and 65, which we call *AgeSpline*, to use in the imputation models.

We specified the following imputation models:

- Female was missing for only one observation (0.01 percent). The imputation model for Female was a logistic regression on AgeSpline, PAY1, and White. Note that we imputed Female for only one observation. We could have simply dropped this one observation without affecting the results. However, because we were imputing other missing values, there was no reason to omit it.

- PAY1 was missing for 30 observations (0.16 percent). The imputation model for PAY1 (six payer categories) was a discriminant function based on Female, AgeSpline, and White because this is the only method available in SAS PROC MI. This method assumes that the predictor variables are distributed as multivariate normal. However, we were imputing PAY1 for only 30 observations in this analysis, so departures from the multivariate normal assumption were not likely to have much effect on the results.

- White (race) was missing for 2,424 observations (13.04 percent). We had difficulty imputing race with all its categories using the discriminant function method, which is the only method available for nominal variables in SAS PROC MI.[7] Therefore, for this exercise, we created a binary variable: White=1 if race was White; White=0 if race was not White and not missing; White=missing if race was missing. The imputation model for White was a logistic regression on AgeSpline, Female, and PAY1.

- Total charges were missing for 4,273 observations (22.99 percent). The imputation model for total charges was a regression of log (total charges) on Female, White, AgeSpline, log (LOS), log (Nchronic+1), log (DaysCCU+1), and log (DaysICU+1). The software back-transformed the predictions of log (total charges) to produce imputations on the original scale (total charges). Total charges were log-transformed for three reasons. First, log (total charges) satisfied the assumptions of OLS regression better than a regression using total charges as the dependent variable. Second, plots based on the nonmissing data indicated better linear correlation between log (total charges) and the predictors than between untransformed total charges and the predictors. Third, this transformation ensured that the imputed values for total charges would always take positive values. Regressions using total charges as the dependent variable produced negative imputations for some observations.

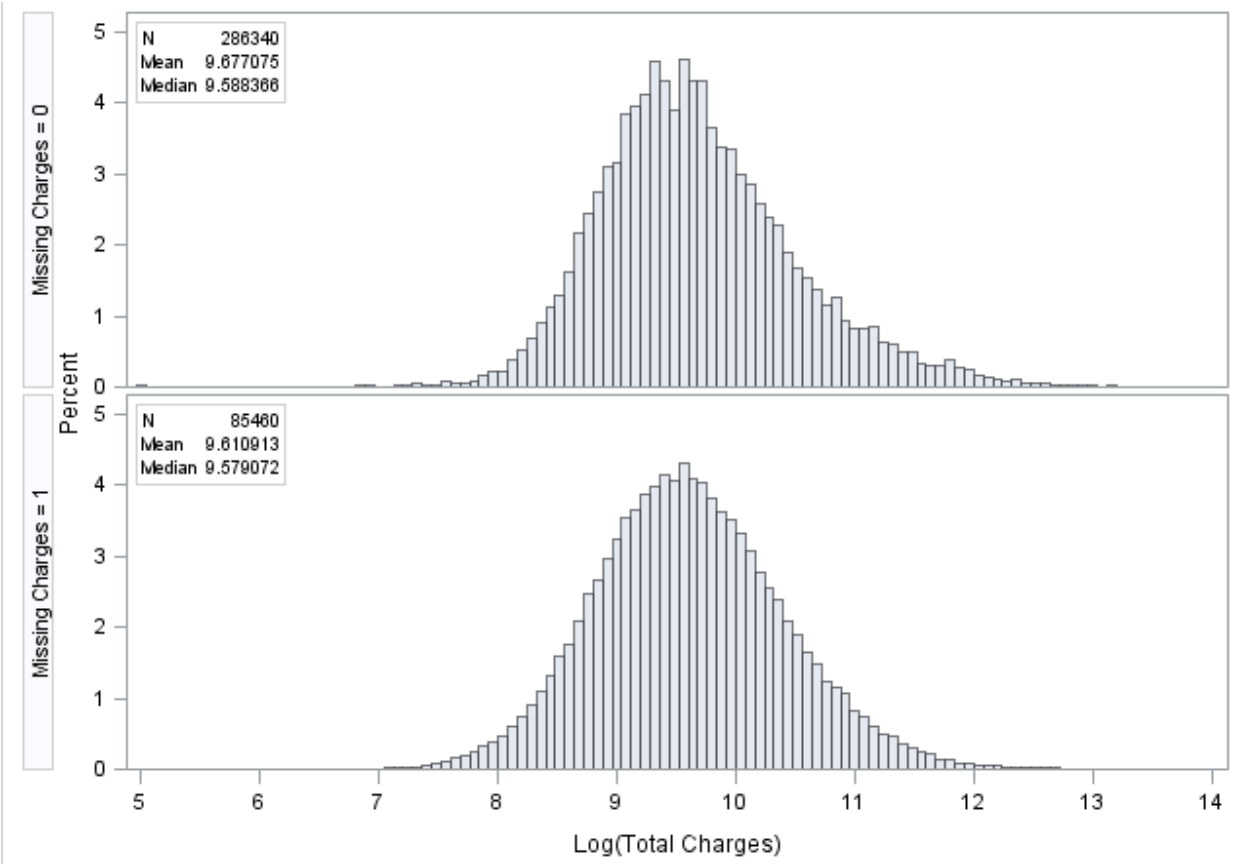
We checked the reasonableness of these imputations by comparing the distributions of the observed values with the distributions of the imputed values. For sex, 48 percent of the observed values were female compared with 40 percent of the imputed values. Given that this was based on 20 imputed values for a single observation, the imputed values appear reasonable for sex. Likewise, 59 percent of the observed values were White for race compared with 60 percent of the imputed values. So the imputations for White race seem to be in line with expectations.

Primary payer was missing for only 30 observations (0.16 percent). The observed percentage for Medicare was 45 compared with 21 for the imputed percentages. This difference was most likely attributable to the fact that observations missing payer were 13 years younger, on average, than were observations not missing payer. Thus, it makes sense that the observations missing payer would be less likely to have Medicare insurance.

[7] SAS Institute suggests potential solutions for this issue (http://support.sas.com/kb/51/472.html). However, the solutions were unsuccessful with our data.

Figure 1 compares the distributions of observed and imputed log (total charges). The top histogram shows the observed distribution, and the bottom histogram shows the imputed distribution. The imputed distribution appears to be sensible compared with the observed distribution.

Figure 1. Distribution of observed and imputed log (charges) for complicated diabetes, Michigan SID 2012



Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), State Inpatient Databases (SID) for Michigan, 2012

Having satisfied ourselves that the imputed values were reasonable, we estimated mean total charges once solely on the basis of the observed data and once on the basis of the data completed with the imputations.

On the basis of the pooled imputed data, the mean total charge was estimated as $23,132 compared with $23,784 solely on the basis of the nonmissing values for total charges—a difference of $652 (2.7 percent). The standard error estimated from the imputed data was $226, which was $37 less than the estimate of $263 based on the nonmissing total charges, presumably because of the increased sample size afforded by the imputations. The 95 percent confidence interval based on the pooled imputed data was ($22,690, $23,574) compared with ($23,268, $24,300) based solely on the nonmissing values for total charges, resulting in only a 19 percent overlap between the endpoints of the two intervals. Therefore, the range of plausible values for mean charges differed substantially between the full data including imputations and the subset of data without missing values.

Other strategies are possible for imputing the missing values in these data. For example, the analyst could perform imputations in two steps: (1) impute only enough missing data to create monotone missing data patterns and (2) use monotone imputation methods to impute the

remaining missing data (Berglund and Heeringa, 2014). Other imputation models are also possible. For instance, rather than impute from a regression of log-transformed total charges, we could have imputed from a regression of untransformed total charges and then used predictive mean matching, which would have ensured imputed values that stayed within the range of the nonmissing values for total charges. In fact, analysts are encouraged to perform sensitivity analyses by applying more than one missing value method and assessing the differences (e.g., Van Buuren, 2012).

Also, the number of variables included in this analysis was restricted for simplicity of exposition. However, other variables could be included that might make the imputations more accurate. For example, if the missingness of charges was related to whether patients were treated surgically, then it would be beneficial to include surgery indicators in the imputation model. Also, hospital indicators or hospital characteristics could be included in the imputation model to account for hospital-specific effects. Although hierarchical models seems to be an appropriate choice for clustered data (patients clustered within hospitals), imputation based on hierarchical models has not been fully developed as of this writing (Van Buuren, 2012).

6.1.2 SID Regression of Total Charges for Complicated Diabetes

For this example, we regressed (the logarithm of) total charges on age, sex, race, primary payer, LOS, number of chronic conditions, CCU days, and ICU days. We used the M=20 data sets imputed in the last section using the FCS method and compared the pooled estimates of the regression coefficients and their significance levels with a regression based solely on the 67 percent of cases with nonmissing data for all of the data elements.

For illustrative purposes, we fit a regression using PROC GLIMMIX with hospitals as random effects with the assumption that total charges are distributed as log-normal. Although this model does not have any hospital-level predictors, they could be added to help explain hospital-level variation. We fit the model once using the original data, deleting all observations with missing values (listwise deletion). We then fit the model to each of the 20 imputations produced in the previous section and pooled the results.

Table 5 compares the coefficients, standard errors, and p-values between the original estimates and the pooled imputation estimates. The parameter names are self-explanatory with the exception of the age parameters, which are spline coefficients. This is a log-linear model; therefore the antilogarithm of a coefficient estimates the *multiplicative* effect of a one-unit increase in the parameter.

Table 5. Pooled versus original regression coefficients for complicated diabetes, Michigan SID 2012

Parameter	Original coefficient	Pooled coefficient	Original std error	Pooled std error	Original p-value	Pooled p-value
Intercept	7.258	7.415	0.1876	0.1970	<.0001	<.0001
Female	–0.033	–0.025	0.0073	0.0068	<.0001	0.0003
White	0.045	0.030	0.0091	0.0090	<.0001	0.0012
Age_0	–0.070	0.028	0.2106	0.2154	0.7389	0.8971
Age_1	0.426	0.402	0.1769	0.1861	0.0160	0.0327
Age_2	0.322	0.190	0.1987	0.2084	0.1054	0.3644
Age_3	0.515	0.435	0.1645	0.1762	0.0018	0.0152
Age_4	0.147	0.004	0.2236	0.2351	0.5116	0.9866
Medicaid	–0.050	–0.068	0.0118	0.0111	<0.0001	<0.0001
Private	–0.018	–0.028	0.0105	0.0106	0.0861	0.0089
Self Pay	–0.012	–0.039	0.0170	0.0165	0.4941	0.0187
No Charge	–0.077	–0.090	0.0420	0.0427	0.0675	0.0357
Other Pay	–0.047	–0.022	0.0290	0.0263	0.1059	0.4055
Log(LOS+1)	0.995	0.991	0.0070	0.0062	<0.0001	<0.0001
Log(Nchronic+1)	0.166	0.165	0.0098	0.0093	<0.0001	<0.0001
Log(CCU Days+1)	0.220	0.220	0.0182	0.0187	<0.0001	<0.0001
Log(ICU Days+1)	0.278	0.269	0.0105	0.0109	<0.0001	<0.0001

Abbreviations: LOS, length of stay; CCU, coronary care unit; ICU, intensive care unit.

Note: Generalized linear model fit using SAS PROC GLIMMIX under the assumption that total charges are log-normally distributed. Consequently, the regression predicts log (total charges).

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), State Inpatient Databases (SID) for Michigan, 2012

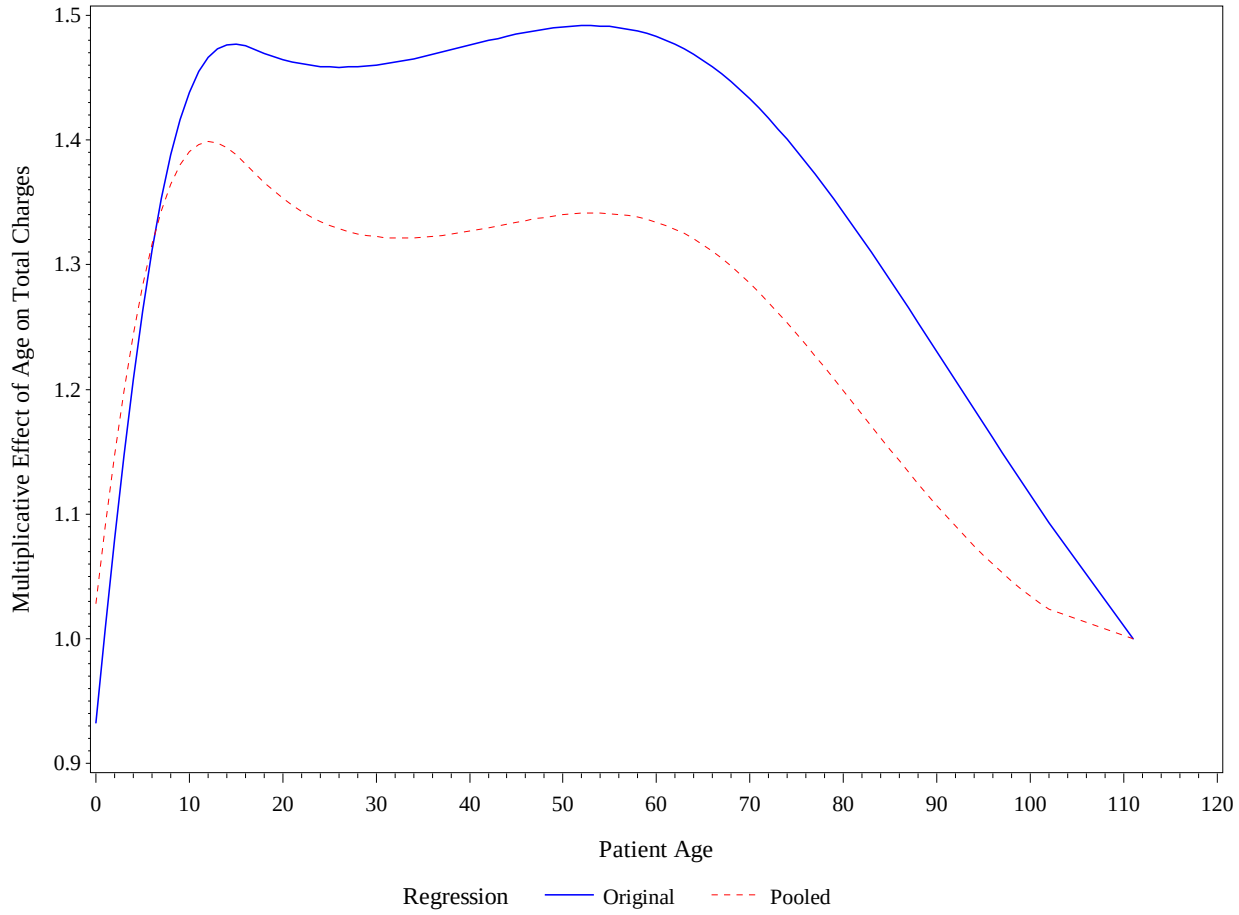
For example, on the basis of the original regression estimates, total charges were approximately 3.3 percent lower for females than for males, holding everything else constant. On the basis of the pooled regression estimates, they were about 2.5 percent lower. Medicare was the reference primary payer category (omitted). Consequently, on the basis of the original regression, charges for Medicaid patients were about 5 percent lower than charges for Medicare patients. On the basis of the pooled regression, charges for Medicaid patients were about 6.8 percent lower. At the 5 percent significance level, the original regression charges for patients of the remaining payers were not statistically significant. However, except for OtherPay, they were significant at the 5 percent level on the basis of the pooled regression. Thus, imputing missing values may alter the conclusions drawn from the regression.

The estimates for the remaining parameters were not very different between the original and the pooled regression. It is interesting that the coefficient on log (LOS+1) was nearly equal to 1.0.

Thus, the estimated effect of LOS was to multiply total charges by (LOS+1). The estimated multiplicative effects of age are shown in

Figure 2. The original regression is represented by the red line, and the pooled regression is represented by the blue line. The difference between the levels of the two lines is not important (the intercepts differ between the two regressions). However, the relative age effects were mildly different over the age range.

Figure 2. Effect of age on total charges for complicated diabetes, original versus pooled regressions, Michigan SID 2012



Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), State Inpatient Databases (SID) for Michigan, 2012

Although the pooled estimates are subject to the usual caveats that apply to any statistical analysis—we assume that the regression model is correct, that the data are MAR, and that the imputation models are correct—the main reason for assuming that the pooled estimates are preferable to the original estimates is that it is untenable to assume that the data are missing completely at random. In this particular example, we suspect that the charges are MAR (missingness depends on other covariates) and we cannot easily argue that they are MCAR.

6.2 NIS Examples

In this section, we used the 2012 NIS data. In section 3.1, we discussed and provided an overview of missing value rates for selected NIS data elements. For both examples, we imputed missing values for age, sex, race, primary payer, LOS, and total charges. The first example represents a descriptive analysis, and the second example represents a more sophisticated regression analysis.

Although similar to the corresponding analyses of the Michigan SID data, these analyses of the NIS are slightly more complex because the NIS is based on a sample design. This design and the sample weights (Houchens and Elixhauser, 2005) usually must be incorporated when estimates are generated for each of the imputations. Moreover, the degrees of freedom for the estimates from the individual imputations must be passed to SAS PROC MIANALYZE so that it can properly calculate pooled estimates, their standard errors, and confidence intervals. Appendix B contains the SAS code for these examples.

For earlier years of the NIS, the sample design included strata, clusters, and *unequal* sample weights. Consequently, to the extent possible, one should take those design elements into account when imputing missing data in the NIS prior to 2012. For further details, examples, and recommendations, see Berglund and Heeringa (2014) and Heeringa, West, and Burglund (2010). Kim et al. (2006) found that multiple imputation can produce biased results when the sample weights are unequal. However, this concern is likely to be mitigated by the low variance among the NIS sample weights, as long as the weights and other sample design elements are incorporated into the imputation procedure.[8] An alternative for earlier years of the NIS, not used in this example, is to use single imputation and calculate variances according to the procedures described by Wang and Robins (1998) and Robins and Wang (2000), which require special programming.

6.2.1 NIS Average Charges for Complicated Diabetes

Our goal was to estimate the mean and a 95 percent confidence interval for total charges for discharges with complicated diabetes. We used age, sex, race, LOS, number of chronic conditions, performance of a major procedure, primary payer, and NIS sampling stratum indicators to impute the 2.1 percent of discharges with missing values for total charges. Age, sex, race, LOS, and primary payer were also missing for some observations. Therefore, we also needed to impute missing values for those data elements.

The inclusion of sample design elements, such as sampling strata and weights, are normally taken into account when one produces estimates from the NIS (Houchens and Elixhauser, 2005). However, they are also recommended as predictors in the imputation models (Berglund and Heeringa, 2014). The 2012 NIS is a stratified, self-weighted sample. Therefore, we used stratum indicators in the imputation models, but there was no need to use sample weights as

[8] For example, the overall 2011 NIS discharge weights averaged 4.8, with a standard deviation of 0.6, and ranged in value from 3.1 to 24.8. Further, the weights were constant within 55 of the 61 sample strata.

predictors in the imputation models, because the weights were equal across all observations in the sample.

Table 6 shows the missing data pattern in the NIS 2012 for the 105,606 sample discharges with complicated diabetes identified using AHRQ's Clinical Classification Software (DXCCS1=50). The missing values produced 14 distinct missing value groups, numbered in the first column. There are nine columns representing nine data elements corresponding to the NIS sample stratum (NIS_STRATUM), number of chronic conditions (NCHRONIC), major operating room procedure (ORPROC), sex (FEMALE), LOS, age, primary payer (PAY1), total charges (TOTCHG), and race. An "X" means that the data element was not missing in the corresponding group. A "." means that the data element was missing. The final two columns contain the frequency (N) and percentage of cases for each missing value group. For the noncategorical data elements, the means of the nonmissing values are shown in Table 7.

Table 6. Missing data patterns for complicated diabetes, NIS 2012

Group	NIS STRATUM	NCHRONIC	ORPROC	FEMALE	LOS	AGE	PAY1	TOTCHG	RACE	N	%
1	X	X	X	X	X	X	X	X	X	98,701	93.46
2	X	X	X	X	X	X	X	X	.	4,374	4.14
3	X	X	X	X	X	X	X	.	X	2,139	2.03
4	X	X	X	X	X	X	X	.	.	71	0.07
5	X	X	X	X	X	X	.	X	X	240	0.23
6	X	X	X	X	X	X	.	X	.	34	0.03
7	X	X	X	X	X	X	.	.	X	1	0.00
8	X	X	X	X	X	.	X	X	X	2	0.00
9	X	X	X	X	.	X	X	X	X	2	0.00
10	X	X	X	X	.	.	X	X	X	34	0.03
11	X	X	X	.	X	X	X	X	X	2	0.00
12	X	X	X	.	X	X	X	X	.	1	0.00
13	X	X	X	.	X	.	X	X	X	2	0.00
14	X	X	X	.	X	.	.	X	.	3	0.00
Total missing (%)	0 (0.00)	0 (0.00)	0 (0.00)	8 (0.01)	36 (0.03)	41 (0.04)	278 (0.26)	2,211 (2.09)	4,483 (4.25)	6,905 (6.54)	

Abbreviations: LOS, length of stay.

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), National Inpatient Sample (NIS), 2012

Group 1 is the group of discharges with complete data on all nine data elements, representing 98,701 (93.46 percent) of the 105,606 discharges. Thus, about 6.5 percent of the discharges were missing at least one of the nine data elements. NIS stratum, chronic conditions, and major OR procedure were not missing for any of the complicated diabetes discharges. Female was

missing for only eight discharges (groups 11–14), LOS was missing for 36 discharges (groups 9 and 10), age was missing for 41 discharges (groups 8, 10, 13, and 14), and primary payer was missing for 278 discharges (groups 5–7 and 14). Although we could discard the 319 observations (0.30%) missing these four data elements, there is no good argument for excluding them because we can impute them easily alongside the imputations for race and total charges. If these were the only four data elements with missing values, then we could consider omitting them from the analyses if it could be argued that their effects would be trivial.

Table 7. Missing data patterns for complicated diabetes, group means, NIS 2012

Group	NIS STRATUM	NCHRONIC	ORPROC	FEMALE	LOS	AGE	PAY1	TOTCHG	RACE	N	%
1	-----	6.14	0.20	0.47	4.63	51.40	-----	34,343	-----	98,701	93.46
2	-----	5.96	0.19	0.45	4.33	48.71	-----	24,498	.	4,374	4.14
3	-----	7.64	0.25	0.46	4.04	52.61	-----	.	-----	2,139	2.03
4	-----	7.27	0.31	0.39	4.55	52.20	-----	.	.	71	0.07
5	-----	4.81	0.16	0.49	3.80	48.00	.	26,428	-----	240	0.23
6	-----	5.71	0.24	0.38	3.50	44.29	.	16,178	.	34	0.03
7	-----	1.00	0.00	1.00	1.00	16.00	.	.	-----	1	0.00
8	-----	10.00	0.50	0.00	4.50	.	-----	27,861	-----	2	0.00
9	-----	8.50	0.00	0.50	.	83.00	-----	1,142,946	-----	2	0.00
10	-----	2.91	0.00	0.32	.	.	-----	10,460	-----	34	0.03
11	-----	2.00	0.00	.	2.00	20.50	-----	5,678	-----	2	0.00
12	-----	4.00	0.00	.	1.00	56.00	-----	1,193	.	1	0.00
13	-----	2.00	0.00	.	1.50	.	-----	12,878	-----	2	0.00
14	-----	3.00	0.00	.	2.67	.	.	19,237	.	3	0.00

Abbreviations: LOS, length of stay. Dashes indicate non-missing nominal variables for which averages are meaningless.

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), National Inpatient Sample (NIS), 2012

The missing value rates were higher for race (2.1 percent missing overall) and total charges (4.3 percent missing overall). Note also, for example, in Table 7 that the observed mean LOS varied among the first three groups. The observed mean total charge for group 2 was higher than the observed mean charge for group 1, mirroring the differences in the observed mean LOS between the two groups. This result suggests potentially important differences in total charges between cases with and without missing race, and it indicates that LOS might be especially helpful for imputing missing values for total charges.

The missing value pattern is not monotone (Table 6). For example, groups 2 and 3 indicate that some observations have missing charges but not missing race, and other observations have missing race but not missing charges. The missing values were for a mixture of binary, categorical, and continuous variables. We found that the effect of age often had inflections at age 18 and age 65. Regression splines are designed to fit flexible, continuous nonlinear

functions for regression predictors. Therefore, we created a cubic spline function for age with knots at 18 and 65, which we call *AgeSpline*, to use in the regression models.

We took a two-step approach. Age, sex, and LOS were missing for only 46 observations. So, for the first step, we decided to use the MCMC method to impute missing values for these three data elements for its relative simplicity. The MCMC method models the correlation among the three data elements, but it does not model them as a function of any other predictors. For the second step, we then used the FCS method to impute missing values for race, primary payer, and total charges using the previously imputed values for age, sex, and LOS, along with other predictors.

For the first step, we imputed age, sex, and LOS using the MCMC method for multivariate imputation, which assumes that the three data elements are multivariate normal. The distribution of age was fairly symmetric, so we did not transform it. Sex is binary, but we were imputing missing sex values for only eight observations, so we were not concerned about the lack of a normal distribution for that data element. LOS is a nonnegative integer-valued data element with a skewed distribution. Therefore, we modeled log (LOS+1), which was reasonably symmetric.

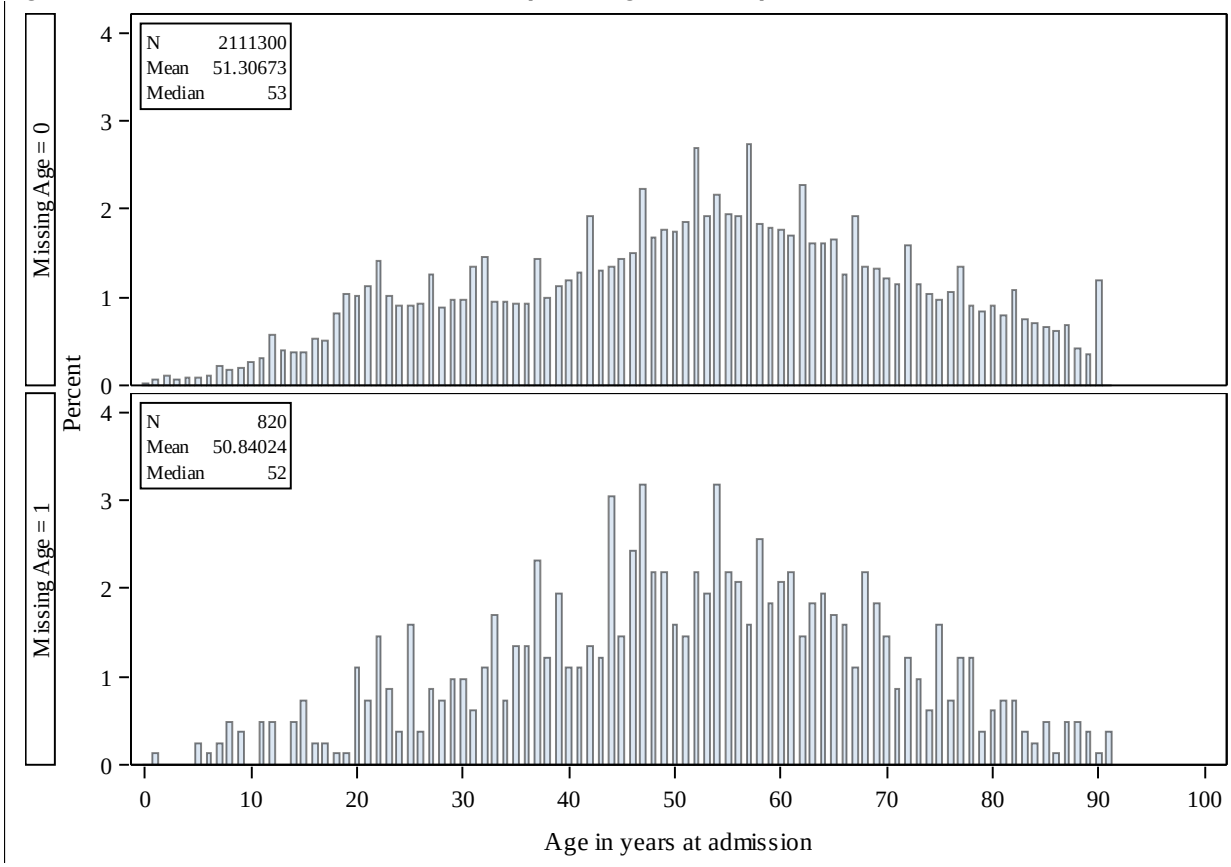
Consistent with the assumption of multivariate normality, the MCMC method produces estimates that theoretically vary continuously from negative infinity to positive infinity. To impute missing values for sex, we mapped the MCMC estimates to values of 0 (for male) and 1 (for female). To impute missing values for LOS, we applied the antilogarithm to the MCMC estimate for log (LOS+1), rounded the result to the nearest integer, constrained values to lie between 1 and 361, and subtracted 1 from the result. Finally, we imputed age by rounding the MCMC estimate for age to the nearest integer, and then by imposing a lower bound of 1 and an upper bound of 90 for age.[9] We produced M=20 imputations.

We checked the reasonableness of these imputations by comparing the distributions of the observed values to the distributions of the imputed values. For sex, 47 percent of the observed values were female compared with 50 percent of the imputed values. Thus, the imputed values appeared reasonable for sex.

Figure 3 and Figure 4 show histograms comparing the observed with the imputed distributions for age and log (LOS+1), respectively. In each Figure, the top histogram shows the distribution of observed values and the bottom histogram shows the distribution of imputed values. The histogram sample size (N) was 20 times the original sample size because the imputed data comprised 20 complete copies of the observations. Considering that we imputed age for only 41 observations (820 imputations) and we imputed LOS for only 36 observations (720 imputations), the distributions of observed and imputed values appear quite similar for both age and log (LOS+1). Thus, we concluded that these imputed values were also reasonable.

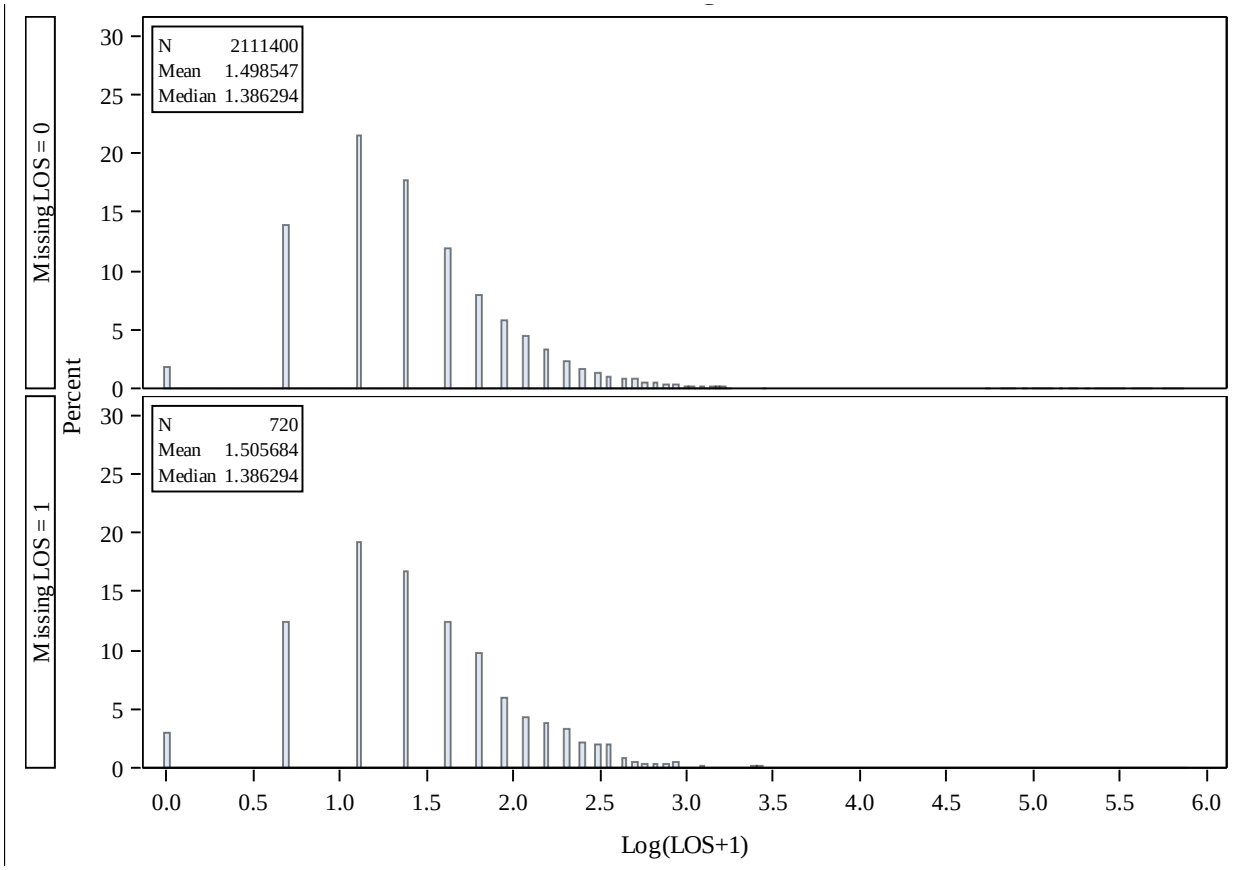
[9] To protect patient confidentiality, age is capped at 90 in the NIS, so ages above 90 are recoded to 90.

Figure 3. Distribution of observed and imputed age for complicated diabetes, NIS 2012



Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), National Inpatient Sample (NIS), 2012

Figure 4. Distribution of observed and imputed log (LOS+1) for complicated diabetes, NIS 2012



Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), National Inpatient Sample (NIS), 2012

For the second step, we used the FCS method for multivariate imputation of race, primary payer (PAY1), and total charges. We specified the following imputation models for these three data elements:

- White (race) was missing for 4,483 observations (4.25 percent). We had difficulty imputing race with all its categories using the Discriminant function method, which is the only method available for nominal variables in SAS PROC MI. Therefore, for this exercise, we created a binary variable: White=1 if race was White; White=0 if race was not White and not missing; White=missing if race was missing. The imputation model for White was a logistic regression on AgeSpline, FEMALE, PAY1, and NIS_STRATUM.
- Primary payer (PAY1) was missing for 278 observations (0.26 percent). The imputation model for PAY1 (six payer categories) was a discriminant function based only on age. Models involving indicators for female, race, and NIS strata failed, probably because this method assumes that the predictor variables are distributed as multivariate normal. However, we were imputing PAY1 for only 0.26 percent of the observations, so the omission of other potential predictors was unlikely to affect the results.

- Total charges were missing for 2,211 observations (2.09 percent). The imputation model for total charges was a regression of log (total charges) on Female, White, AgeSpline, log (LOS), log (Nchronic+1), ORPROC, and NIS_STRATUM. The software back-transformed the predictions of log (total charges) to produce imputations on the original scale (total charges). Total charges were log-transformed for three reasons. First, log (total charges) better satisfied the assumptions of OLS regression compared with a regression using total charges as the dependent variable. Second, plots based on the nonmissing data (not shown) indicated better linear correlation between log (total charges) and the predictors than between untransformed total charges and the predictors. Third, this transformation ensured that the imputed values for total charges would always take positive values.

We compared the distribution of observed values with the distribution of imputed values for White, PAY1, and log (total charges).

Among the observed race values, 55 percent were White. Among the imputed race values, 64 percent were White. This difference might be attributable to a difference in the primary payer distribution between observations missing race and not missing race. Race was missing relatively more often for private payers, and White patients were relatively more likely to have private insurance. Thus, the difference in the percentage of White patients between the observed and imputed race values is not alarming. Indeed, this situation is the very reason why imputation can be so valuable.

The distributions for the primary payer categories are shown in Table 8. The distributions match relatively well considering that only 0.26 percent of the observations were missing primary payer.

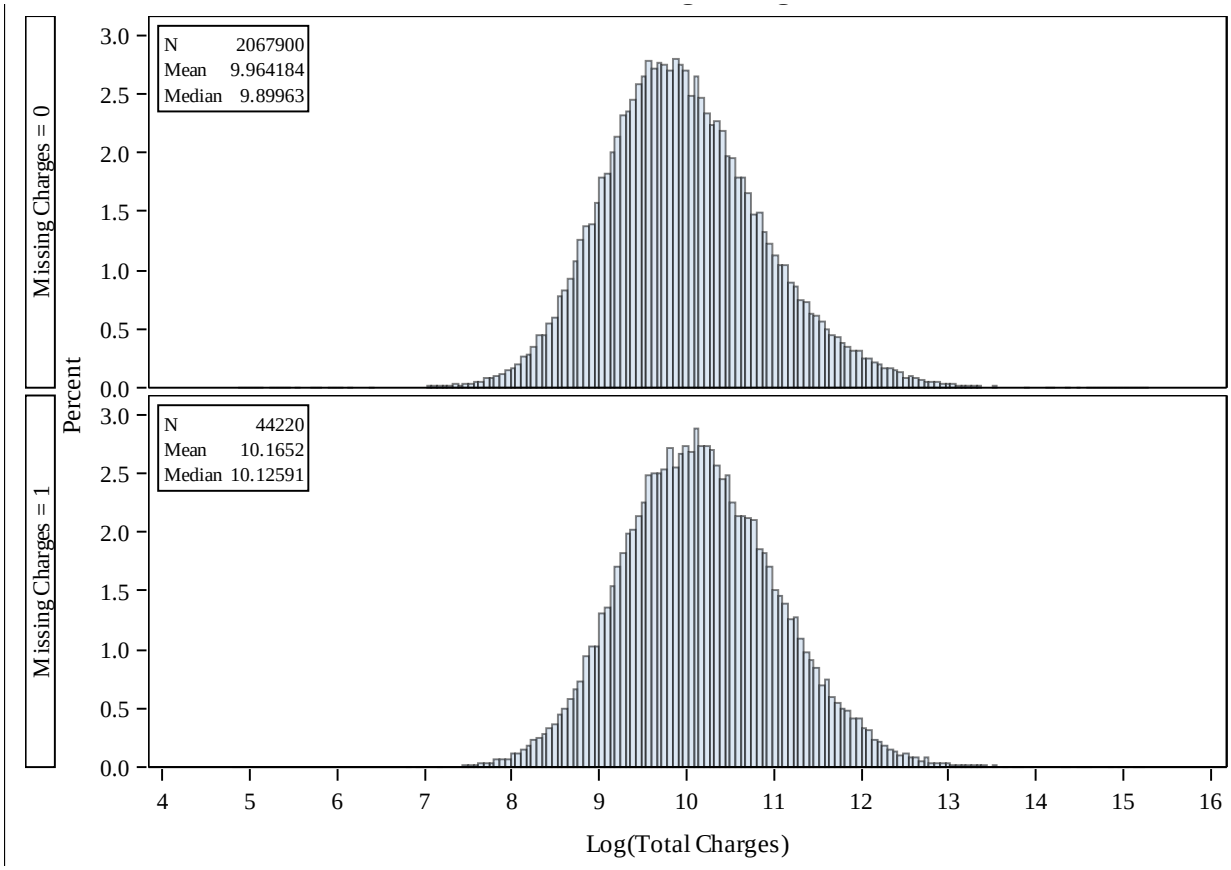
Table 8. Percentage by primary payer and imputation for complicated diabetes, NIS 2012

Primary payer	1 Imputed?	
	No	Yes
Medicare	41.6	35.0
Medicaid	20.5	23.3
Private	21.7	22.8
Self pay	11.4	13.2
No charge	0.9	1.1
Other payer	4.0	4.6

Figure 5 compares the distribution of observed log (charges) to imputed log (charges). The top histogram shows the distribution of observed values, and the bottom histogram shows the

distribution of imputed values. The histogram sample size (N) is 20 times the original sample size because the imputed data comprised 20 complete copies of the observations. The two distributions appear to be quite similar.

Figure 5. Comparison of observed and imputed distribution of log (charges) for complicated diabetes, NIS 2012



Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), National Inpatient Sample (NIS), 2012

For the NIS, it is necessary to account for the sample design elements. Consequently, we used a SAS procedure for survey data, PROC SURVEYMEANS, to estimate means and standard errors both from the original observed data and from the imputed data.

On the basis of the pooled imputed data, the mean total charge was estimated as $34,041 compared with $33,914 based solely on the nonmissing values for total charges—a difference of only $127 (0.4%). The standard error estimated from the imputed data was $414, which was only $6 less than the estimate of $420 based on the nonmissing total charges. The 95 percent confidence interval based on the pooled imputed data was ($33,229, $34,853) compared with ($33,091, $34,737) based solely on the nonmissing values for total charges. This resulted in an 86 percent overlap between the endpoints of the two intervals. Therefore, the range of plausible values for mean charges did not seem to differ substantially between the imputed and nonimputed data.

6.2.2 NIS Regression of Total Charges for Complicated Diabetes

Here we regressed total charges on age, sex, race, primary payer, LOS, number of chronic conditions, a major surgery indicator, and NIS stratum indicators. We used the M=20 data sets imputed in the last section using the MCMC method and the FCS method and compared the pooled estimates of the regression coefficients and their significance levels with a regression based solely on the 93.5 percent of cases with nonmissing data for all of the data elements.

We fit a regression using SAS PROC GLIMMIX with hospitals as random effects with the assumption that total charges are distributed as log-normal. Although this model does not have any hospital-level predictors, they could be added to help explain hospital-level variation. We fit the model once using the original data, deleting all observations with missing values (listwise deletion). We then fit the model to each of the 20 imputations produced in the previous section and pooled the results.

We could have used PROC SURVEYREG to account for the NIS sample design. However, it is acceptable to use other (nonsurvey) regression procedures as long as the NIS design elements are incorporated, which we do by including the NIS stratum indicators as predictors.

Table 9 compares the coefficients, standard errors, and p-values between the original estimates and the pooled imputation estimates. The parameter names are self-explanatory with the exception of the age parameters, which are spline coefficients. The coefficients on the NIS stratum indicators are of no interest, so they are omitted from the table. This is a log-linear model; therefore the antilogarithm of a coefficient estimates the *multiplicative* effect of a one-unit increase in the parameter.

For example, on the basis of the original regression estimates, total charges were approximately 1.6 percent higher for females than for males, holding everything else constant. On the basis of the pooled regression estimates, they were about 1.7 percent higher. Medicare was the reference primary payer category (omitted). Consequently, on the basis of the original regression charges for Medicaid patients were about 3.7 percent lower than charges for Medicare patients. On the basis of the pooled regression, charges for Medicaid patients were about 3.6 percent lower. All of the coefficients are significant at the 5 percent significance level for both sets of estimates.

The multiplicative effects of age are compared in Figure 6. The difference in the level of the two curves is not important because the regressions have slightly different intercepts. The relative age effects appear to be nearly the same between the original and the pooled regressions.

The parameters estimates were not very different between the original and the pooled regressions. Consequently, inferences were about the same whether they were based on an analysis that deletes all observations with missing values or they were based on an analysis that imputes missing values. Nevertheless, we had to perform the analysis before we could come to that conclusion. If the original regression results are presented, the analyst can confidently say that the omission of cases with missing values had no discernible effect on the results. In any case, the pooled estimates are subject to the usual caveats that apply to any

statistical analysis—we assume that the regression model is correct, that the data are MAR, and that the imputation models are correct.

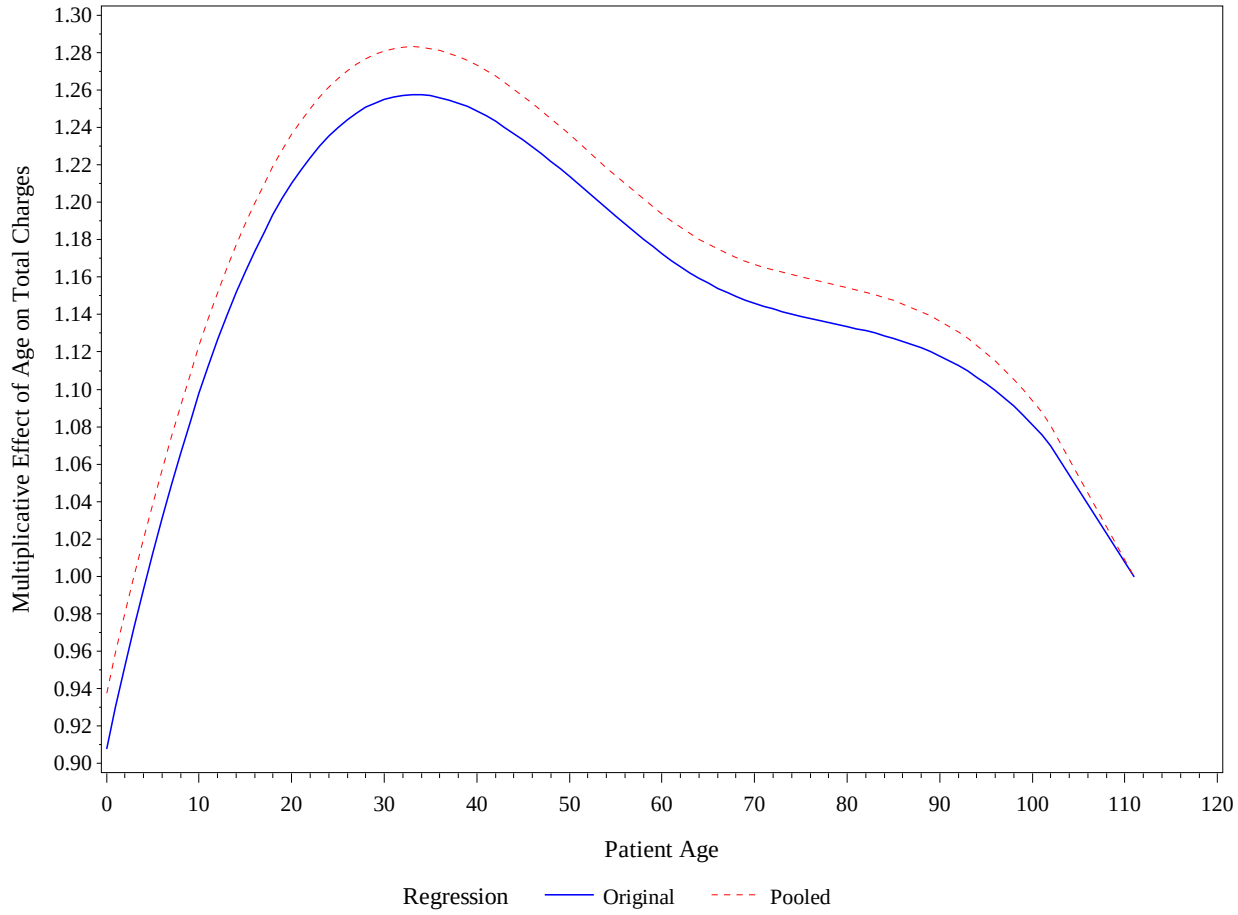
Table 9. Pooled versus original regression coefficients for complicated diabetes, NIS 2012

Parameter	Original coefficient	Pooled coefficient	Original std error	Pooled std error	Original p-value	Pooled p-value
Intercept	8.688	8.668	0.3852	0.3819	<0.0001	<0.0001
Female	0.016	0.017	0.0025	0.0025	<0.0001	<0.0001
White	0.024	0.021	0.0030	0.0030	<0.0001	<0.0001
Age_0	-0.097	-0.064	0.0325	0.0319	0.0030	0.0434
Age_1	0.053	0.074	0.0149	0.0153	0.0004	<0.0001
Age_2	0.338	0.360	0.0147	0.0160	<0.0001	<0.0001
Age_3	0.071	0.086	0.0104	0.0109	<0.0001	<0.0001
Age_4	0.149	0.170	0.0128	0.0148	<0.0001	<0.0001
Medicaid	-0.037	-0.036	0.0040	0.0040	<0.0001	<0.0001
Private	-0.014	-0.015	0.0038	0.0038	0.0002	<0.0001
Self Pay	-0.023	-0.022	0.0048	0.0048	<0.0001	<0.0001
No Charge	-0.065	-0.065	0.0140	0.0140	<0.0001	<0.0001
Other Pay	-0.025	-0.023	0.0072	0.0071	0.0005	0.0011
Log(LOS)	0.903	0.903	0.0024	0.0024	<0.0001	<0.0001
Log(Nchronic)	0.113	0.114	0.0034	0.0034	<0.0001	<0.0001
OR Proc	0.378	0.379	0.0036	0.0036	<0.0001	<0.0001

Note: Generalized linear model fit using SAS PROC GLIMMIX under the assumption that total charges are log-normally distributed. Consequently, the regression predicts log (total charges).

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), National Inpatient Sample (NIS), 2012

Figure 6. Multiplicative effect of age on total charges for complicated diabetes, original versus pooled regression, NIS 2012



Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project (HCUP), National Inpatient Sample (NIS), 2012

7. RECOMMENDATIONS

We end with some general recommendations concerning the handling of missing values in HCUP data:

1. Do not ignore missing values! It is important to acknowledge missing data and to think about the implications for your analysis, regardless of the missing value rate.
2. Use the HCUP Web site and any other relevant information sources to research how missing values are coded and why they occur for certain data elements. This step is a key to understanding the missing value mechanism and to specifying imputation models.
3. Compare characteristics between observations with and without missing values. Do the observations with missing values tend to be different from observations without missing values?
4. If the missing data rate is very low and a convincing argument can be made that the data are missing completely at random, then the observations with missing values can be safely dropped from the study. If there is any doubt, plan to do multiple imputations to determine whether it makes a difference to your analysis.
5. Are you doing univariate imputations (imputing a single variable) or multivariate imputations (imputing more than one variable)? For a given data element, imputation techniques may differ between univariate and multivariate imputations.
6. For multivariate imputations decide on a method: Markov Chain Monte Carlo (MCMC), Full Conditional Specification (FCS), or Monotone. If the missing value pattern is arbitrary but nearly monotone, then consider a two-step process, imputing just enough data to achieve monotonicity at the first step and imputing the remaining data at the second step.
7. If necessary, specify an appropriate imputation model for each variable for which missing values are to be imputed. The methods listed in section 4.5 are a place to start. Be sure that the method matches the variable type for each variable you are imputing.
8. Pay attention to the statistical distributional assumptions. Ordinary least squares is robust to departures from normality, but it is also important to think about other assumptions, such as homoscedastic variances. Perhaps a variable can be transformed to better match a method's assumptions or to ensure that plausible imputations are generated (e.g., positive values when imputing length of stay).
9. It is better to use too many rather than too few predictor variables in the imputation models. This helps to ensure that the imputations yield valid inferences under the MAR assumption or, if the data are actually MNAR, to bring the inferences closer to those under MAR. For example, perhaps an observed variable is correlated with an unobserved variable that governs the missingness mechanism.
10. For multiple imputation, the number of imputations (M) should be large enough to ensure that the pooled estimates are stable. Often 5 to 10 imputations are sufficient. However, more imputations may be required, especially if the missing data rates are high. One strategy is to develop models using 5 to 10 imputations and then increase the number of imputations for the final models.
11. Check that the imputed values seem reasonable. For example, compare the distribution of observed values with the distribution of imputed values. If the imputed values are clearly "wrong," then look for errors in your code or consider using another imputation model. If the imputed values are distributed differently from the way that the nonimputed values are distributed, then look to the data for an explanation. For example, perhaps

the observations with missing values have a different age or sex distribution than the observations without missing values. This is where multiple imputation pays dividends because the observations that are missing data are somehow different from the observations that are not missing data.

12. Once you have M data sets, each with a different set of imputations, estimate statistics of interest from each of the M data sets in the same way that you would have produced estimates using the original data had it been free of missing data. Statistics of interest might be descriptive statistics, correlations, or regression coefficients, for example.
13. Estimate the statistics of interest by pooling the M estimates. Most statistical software use Rubin's rules for this.
14. Compare the pooled statistics with the statistics estimated (by default in most statistical packages) solely on the basis of observations without missing values. If the two estimates are very close, then choose one estimate for inferences but report the fact that you tried it both ways and that it did not make a difference.
15. If the pooled statistics are substantially different from the statistics based solely on the nonmissing data, then consider using a different imputation model to test the sensitivity of the pooled statistics to your choice of imputation method.

Finally, seek published examples of missing data problems that are similar to yours. Textbooks on missing data usually offer worked examples from different fields of study, including health services research. Also, a number of journal articles have been written on the topic of missing data, as well as studies that use missing value methods in their analysis.

8. REFERENCES

Allison PD. Imputation of categorical variables with PROC MI. In: Proceedings of the Thirtieth Annual SAS® Users Group International Conference. Cary, NC: SAS Institute; 2005:113-30.

Andridge RR, Little RJA. A review of hot deck imputation for survey non-response. International Statistical Review 2010;78(1):40-64.

Aycock AC. A Gentle Introduction to Stata, 4th ed. College Station, TX: Stata Press; 2014. http://www.stata.com/bookstore/gentle-introduction-to-stata/. Accessed October 8, 2014.

Berglund P, Heeringa S. Multiple Imputation of Missing Data Using SAS. Cary, NC: SAS Institute; 2014.

Carpenter JR, Kenward MG. Multiple Imputation and its Application. New York: Wiley; 2013.

Cochran WG. Sampling Techniques, 3rd ed. New York: Wiley; 1977.

Congdon P. Bayesian Statistical Modelling, 2nd ed. New York: Wiley; 2006.

Foreman EK. Survey Sampling Principles. New York: Marcel Dekker; 1991.

Gelman A, Carlin JB, Stern HS, et al. Bayesian Data Analysis, 3rd ed. New York: CRC Press; 2014.

Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. New York: Cambridge University Press; 2007.

Heeringa SG, West BT, Berglund A. Applied Survey Data Analysis. London: Chapman and Hall; 2010.

Houchens, R. Inferences With HCUP State Databases Final Report. HCUP Methods Series Report # 2010-05. Online October 12, 2010. Rockville, MD: Agency for Healthcare Research and Quality. http://www.hcup-us.ahrq.gov/reports/methods.jsp. Accessed October 13, 2014.

Houchens R, Elixhauser A. Final Report on Calculating Nationwide Inpatient Sample (NIS) Variances. HCUP Method Series Report # 2003-02. Online June 2005 (revised June 6, 2005). Rockville, MD: Agency for Healthcare Research and Quality. http://www.hcup-us.ahrq.gov/reports/methods/CalculatingNISVariances200106092005.pdf. Accessed October 13, 2014.

Houchens, R, Ross, DN, Elixhauser, A, Jiang J. Nationwide Inpatient Sample Redesign Final Report. Deliverable # 1308.11. Rockville, MD: Agency for Healthcare Research and Quality; 2014. http://www.hcup-us.ahrq.gov/db/nation/nis/reports/NISRedesignFinalReport040914.pdf. Accessed October 13, 2014.

Kim JK, Brick, JM, Fuller, WA, Kalton, G. On the bias of the multiple imputation variance estimator in survey sampling. Journal of the Royal Statistical Society, Series B 2006;68:509-21.

Kim JK, Shao J. Statistical Methods for Handling Incomplete Data, Boca Raton, FL: Chapman and Hall/CRC Press; 2014.

Little RJA, Rubin DB. Statistical Analysis With Missing Data, 2nd ed. New York: Wiley; 2002.

Meng, XL. Multiple imputation inferences with uncongenial sources of input (with discussion). Statistical Science 1994;9:538-73

Meng, XL. Missing data: Dial M for ???. Journal of the American Statistical Association 2000, 95(452):1325-30.

National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials. Washington, DC: The National Academies Press; 2010.

R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2011. http://www.R-project.org/. Accessed October 8, 2014.

Raghunathan, TE, Solenberger, PW, Van Hoewyk, JV. IVEware: Imputation and Variance Estimation Software User Guide. Survey Research Center, Institute for Social Research, University of Michigan; March 2002.

Rao JKN, Shao J. Modified balanced repeated replication for complex survey data. Biometrika 1999; 86(2):403-5.

Robins JM, Wang N. Inference for imputation estimators. Biometrika 2000,87:113-24.

Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.

SAS Institute Inc. SAS/STAT Software, Version 9.4. Cary, NC: SAS Institute Inc.; 2014. http://www.sas.com/. Accessed October 8, 2014.

StataCorp LP. Stata Data Analysis Statistical Software, Release 12. College Station, TX: StataCorp; 2011. http://www.stata.com/. Accessed October 8, 2014.

Su Y, Gelman A, Hill J, Yajima M. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. Journal of Statistical Software, 2011;45(2):1-31. http://www.jstatsoft.org/v45/i02/. Accessed October 8, 2014.

Van Buuren S. Flexible Imputation of Missing Data. New York: Chapman & Hall; 2012.

Von Hippel PT. Should a normal imputation model be modified to impute skewed variables? Sociological Methods and Research. 2013;42(1):105-38.

Wei X. %PROC_R: A SAS Macro that Enables Native R Programming in the Base SAS Environment. Journal of Statistical Software, 2012;46(2):1-13. http://www.jstatsoft.org/v46/c02/paper. Accessed October 8, 2014.

Wang N, Robins JM. Large sample inference in parametric multiple imputation. Biometrika 1998; 85:935-48.

Appendix A: SAS Code for Analysis of Complicated Diabetes, Michigan SID 2012

```
*  Subset data to discharges with complicated diabetes ;

data diabetes;
  length missing_totchg missing_race missing_pay1 missing_female
      medicare medicaid private selfpay nocharge otherpay
      white black hispanic asian native otherrace 3;
  set mi_sid_2012_core;
  * Diabetes Mellitus with Complications ;
  if dxccs1 = 50 ;
  *  Set missing value indicators for later use ;
  if missing(totchg) then missing_totchg = 1; else missing_totchg = 0;
  if missing(race) then missing_race = 1; else missing_race = 0;
  if missing(pay1) then missing_pay1 = 1; else missing_pay1 = 0;
  if missing(female) then missing_female = 1; else missing_female = 0;
  *  Transform variables for regressions ;
  log_los= log(LOS + 1);
  log_totchg = log(totchg) ;
  log_ccu_days = log(daysccu+1) ;
  log_icu_days = log(daysicu+1) ;
  log_nchronic = log(nchronic+1) ;
  *  Create indicators for primary payer ;
  if pay1 > . then do;
    Medicare = 0; Medicaid = 0; Private = 0; SelfPay = 0; NoCharge = 0;
OtherPay = 0;
    if pay1 = 1 then Medicare = 1 ;
      else if pay1 = 2 then Medicaid = 1;
      else if pay1 = 3 then Private = 1;
      else if pay1 = 4 then SelfPay= 1;
      else if pay1 = 5 then NoCharge = 1;
      else if pay1 = 6 then OtherPay = 1;
  end;
  *  Create indicators for race ;
  if race > . then do;
    white = 0; black = 0; hispanic = 0; asian = 0; native = 0; otherrace = 0;
    if race = 1 then white = 1;
    else if race = 2 then black = 1;
    else if race = 3 then hispanic = 1;
    else if race = 4 then asian = 1;
    else if race = 5 then native = 1;
    else if race = 6 then otherrace = 1;
  end;
  keep age died dshospid female pay1 race los log_los totchg log_totchg
nchronic log_nchronic
      daysccu log_ccu_days log_icu_days daysicu missing_totchg missing_race
missing_pay1 missing_female
      medicare medicaid private selfpay nocharge otherpay
      white black hispanic asian native otherrace ;
run;

*  Create age splines with knots at 18 and 65;

proc transreg data=diabetes design;
  model bspl(age / knots=18 65 exknots=0 110);
  output out=age_splines ;
run;
```

```
*  Merge age splines to data ;

data diabetes ;
  merge diabetes age_splines(keep=age_0 age_1 age_2 age_3 age_4 age_5);
run;

*  Examine distibutions of binary and categorical variables ;

proc freq data=diabetes ;
  table female died pay1 race / missing ;
  Title 'Distribution of Categorical Variables, Complicated Diabetes, MI SID
2012';
run;

*  Examine distibutions of continuous and count variables ;

proc univariate data=diabetes ;
  var totchg los age nchronic daysccu daysicu;
  histogram totchg los age nchronic daysccu daysicu;
  Title 'Distribution of Continuous Variables, Complicated Diabetes, MI SID
2012';
run;

*  Output the missing value pattern ;

proc mi data=diabetes nimpute=0;
 var age los nchronic daysccu daysicu female pay1 race totchg ;
 format age los nchronic daysccu daysicu female pay1 race 5.2 totchg
comma8.0;
 title 'Missing Data Patterns, Complicated Diabetes, Michigan SID 2012';
run;

* Plot log(charges) vs age ;

proc summary data=diabetes nway;
  class age;
  var log_totchg ;
  output out=plotage mean= ;
run;

proc gplot data=plotage ;
  plot log_totchg * age ;
  symbol i=sm50 ; * Produce scatterplot smoother ;
 title 'Log(Charges) vs. Age, Complicated Diabetes, Michigan SID 2012';
run;

* Plot log(charges) vs log(los+1) ;

proc summary data=diabetes nway;
  class log_los;
  var log_totchg ;
  output out=plotlos mean= ;
run;

proc gplot data=plotlos ;
  plot log_totchg * log_los ;
```

```
  symbol i=sm50 ; * Produce scatterplot smoother ;
 title 'Log(Charges) vs. Log(LOS+1), Complicated Diabetes, Michigan SID
2012';
run;

* Plot log(charges) vs log(Nchronic+1) ;

proc summary data=diabetes nway;
  class log_nchronic;
  var log_totchg ;
  output out=plotchronic mean= ;
run;

proc gplot data=plotchronic ;
  plot log_totchg * log_nchronic ;
  symbol i=sm50 ; * Produce scatterplot smoother ;
 title 'Log(Charges) vs. Log(Chronic+1), Complicated Diabetes, Michigan SID
2012';
run;

* Plot log(charges) vs log(CCU Days+1) ;

proc summary data=diabetes nway;
  class log_ccu_days ;
  var log_totchg ;
  output out=plotccu mean= ;
run;

proc gplot data=plotccu ;
  plot log_totchg * log_ccu_days ;
  symbol i=sm50 ; * Produce scatterplot smoother ;
 title 'Log(Charges) vs. Log(CCU Days+1), Complicated Diabetes, Michigan SID
2012';
run;

* Plot log(charges) vs log(ICU Days+1) ;

proc summary data=diabetes nway;
  class log_icu_days ;
  var log_totchg ;
  output out=ploticu mean= ;
run;

proc gplot data=ploticu ;
  plot log_totchg * log_icu_days ;
  symbol i=sm50 ; * Produce scatterplot smoother ;
 title 'Log(Charges) vs. Log(ICU Days+1), Complicated Diabetes, Michigan SID
2012';
run;

*  Regress log(totchg) on predictors to examine residuals ;
*  All cases with missing values are automatically dropped ;

ods graphics on;
proc reg data=diabetes plots(maxpoints=20000)=(diagnostics
predictions(X=age));
```

```
  model log_totchg = age_0 age_1 age_2 age_3 age_4 age_5 log_los log_nchronic
log_ccu_days log_icu_days female
      medicaid private selfpay nocharge otherpay
      black hispanic asian native otherrace  ;
  var age ;
  output out=preds p=predicted r=residual ;
  title 'Regression of Total Charges on Predictors, Listwise Deletion,
Complicated Diabetes, MI SID 2012';
run;

*  FCS method for imputation ;
*  Create 20 imputations, use SEED argument so that results can be
reproducted ;

proc mi data=diabetes nimpute=20 seed=19520201 out=imputed ;
 *  Declare class variables ;
 class female pay1 white ;
 *  Transform total charges for regression ;
 transform log(totchg) ;
 *  Imputation model for female ;
 fcs logistic(female=age_0-age_4 pay1 white) ;
 *  Imputation model for primary payer ;
 fcs discrim(pay1=female age_0-age_4 white / classeffects=include) ;
 *  Imputation model for white race ;
 fcs logistic(white=female age_0-age_4 pay1 ) ;
 *  Imputation model for total charges ;
 fcs reg(totchg = female white pay1 age_0-age_4 log_los log_nchronic
log_ccu_days log_icu_days ) ;
 *  Declare analysis variables ;
 var age los log_los age_0-age_4 female daysccu log_ccu_days daysicu
log_icu_days
      nchronic log_nchronic pay1 white totchg ;
 Title 'Impute Missing Values Using FCS, Complicated Diabetes, MI SID 2012';
run;

*  Check the reasonableness of the imputations ;

proc freq data=imputed ;
  table female * missing_female pay1 * missing_pay1 white * missing_race ;
  title 'Compare Distributions of Observed Values to Imputed Values,
Complicated Diabetes, MI SID 2012';
run;

proc means data=imputed n mean median ;
 class missing_pay1 ;
 var age ;
  title 'Compare Age Distribution Between Observed Payer and Imputed Payer,
Complicated Diabetes, MI SID 2012';
run;

*  Calculate log(charges) for comparison ;

data imputed ;
  set imputed ;
  log_totchg = log(totchg);
run;
```

```
proc univariate data=imputed ;
  class missing_totchg ;
  var log_totchg ;
  histogram log_totchg ;
  inset n mean median;
  label log_totchg = 'Log(Total Charges)' missing_totchg = 'Missing Charges';
  title 'Compare Distributions Between Imputed and Non-Imputed Values,
Complicated Diabetes, NIS 2012';
run;

*  Calculate means and standard errors for each imputation ;
proc summary data=imputed nway;
  class _imputation_ ;
  var totchg ;
  output out = imputed_means mean= stderr=totchg_SE ;
run;

*  Obtain pooled estimates.  For this analysis, EDF is 1 less than the number
of cases. ;

proc mianalyze data=imputed_means edf=18589 ;
  modeleffects totchg ;
  stderr totchg_SE ;
  Title 'Pooled Estimates Based on FCS Imputations, Complicated Diabetes, MI
SID 2012';
run;

*  Run regressions on original data (listwise deletion) ;

ods output ParameterEstimates=parms_original ;
proc glimmix data=diabetes noclprint  ;
  class dshospid ;
  model totchg = female white age_0-age_4 medicaid private selfpay nocharge
otherpay
      log_los log_nchronic log_ccu_days log_icu_days
     / dist=lognormal solution ddfm=BW ;
  random intercept / sub=dshospid type=VC ;
  Title 'Original Regression Estimates Based on Data with Missing Values,
Complicated Diabetes, MI SID 2012';
run;

*  Run regressions on imputed data ;

ods output ParameterEstimates=parms_by_imputation covb=covb_by_imputation;
proc glimmix data=imputed noclprint ;
  by _imputation_ ;
  class dshospid pay1;
  model totchg = female white age_0-age_4 medicaid private selfpay nocharge
otherpay
      log_los log_nchronic log_ccu_days log_icu_days
     / dist=lognormal solution covb ;
  random intercept / sub=dshospid ;
  Title 'Imputation-Specific Regression Estimates Based on FCS Imputations,
Complicated Diabetes, MI SID 2012';
run;
```

```
*  Create Pooled Estimates for the regression coefficients and their standard
errors ;

ods output ParameterEstimates=parms_pooled ;
proc mianalyze parms=parms_by_imputation edf=18577
               covb(effectvar=rowcol)=covb_by_imputation;
   modeleffects Intercept female white age_0 age_1 age_2 age_3 age_4
      medicaid private selfpay nocharge otherpay
      log_los log_nchronic log_ccu_days log_icu_days;
  Title 'Pooled Regression Estimates Based on FCS Imputations, Complicated
Diabetes, MI SID 2012';
run;

*  Compare original estimates to imputation-based estimates ;

data parameter_estimates ;
  merge parms_pooled
           parms_original(rename=(Estimate=Original_Estimate
stderr=Original_StdErr
                  tValue=Original_tValue Probt=Original_Probt) drop=DF) ;
run;


proc print data=parameter_estimates label ;
  id Parm ;
  var Original_Estimate Estimate
       Original_StdErr StdErr
       Original_Probt Probt  ;
  label Estimate = 'Pooled Coefficient'
           Original_Estimate = 'Original Coefficient'
           StdErr = 'Pooled Std Error'
           Original_StdErr = 'Original Std Error'
           Probt = 'Pooled p-value'
           Original_Probt = 'Original p-value';
  format Original_Estimate Estimate 8.3 Original_StdErr StdErr 8.4 ;
  title 'Pooled vs. Original Regression Coefficients, Complicated Diabetes,
MI SID 2012';
run;


*  Plot the multiplicative effect of age ;

*  Obtain age spline data ;

proc sort data=diabetes out=age_data_splines(keep=age age_0 age_1 age_2 age_3
age_4) nodupkey ;
 by age ;
run;

*  Put coefficients into a single record ;

proc transpose data=parms_original out=coeffs_original suffix=_orig;
  id effect ;
  var estimate ;
run;

proc transpose data=parms_pooled out=coeffs_pooled suffix=_pool;
```

```
  id parm ;
  var estimate ;
run;

*  Calculate multiplicative effects ;

data plot_age_data ;
  set age_data_splines ;
  if _n_=1 then do;
      set coeffs_original(keep=age_0_orig age_1_orig age_2_orig age_3_orig
age_4_orig) ;
      set coeffs_pooled(keep=age_0_pool age_1_pool age_2_pool age_3_pool
age_4_pool) ;
  end;
  retain age_0_orig age_1_orig age_2_orig age_3_orig age_4_orig
             age_0_pool age_1_pool age_2_pool age_3_pool age_4_pool;
  age_effect_original = exp(age_0 * age_0_orig + age_1 * age_1_orig + age_2 *
  age_2_orig + age_3 * age_3_orig + age_4 * age_4_orig) ;
  age_effect_pooled = exp(age_0 * age_0_pool + age_1 * age_1_pool + age_2 *
  age_2_pool + age_3 * age_3_pool + age_4 * age_4_pool) ;
run;

*  Plot multiplicative effects ;

axis1 label=(angle=90 'Multiplicative Effect of Age on Total Charges') ;
legend1 label=('Regression') value=('Original' 'Pooled');
symbol1 i=join w=2 color=blue line=1;
symbol2 i=join w=2 color=red line=2;
proc gplot data=plot_age_data ;
  plot Age_effect_Original*Age Age_effect_Pooled*Age / overlay vaxis=axis1
  legend=legend1 ;
  label age='Patient Age';
  Title 'Multiplicative Effect of Age, Original vs. Pooled Regression,
Complicated Diabetes, MI SID 2012';
run;
```

Appendix B: SAS Code for Analysis of Complicated Diabetes, NIS 2012

```
*  Subset data to discharges with complicated diabetes ;

data diabetes;
  length missing_totchg missing_race missing_los missing_age missing_pay1
missing_female
      medicare medicaid private selfpay nocharge otherpay
      white black hispanic asian native otherrace 3;
  set nis_2012_core;
  * Diabetes Mellitus with Complications ;
  if dxccs1 = 50 ;
  *  Set missing value indicators for later use ;
  if missing(totchg) then missing_totchg = 1; else missing_totchg = 0;
  if missing(los) then missing_los = 1; else missing_los = 0;
  if missing(race) then missing_race = 1; else missing_race = 0;
  if missing(age) then missing_age = 1; else missing_age = 0;
  if missing(pay1) then missing_pay1 = 1; else missing_pay1 = 0;
  if missing(female) then missing_female = 1; else missing_female = 0;
*  Transform variables for regressions ;
  age_plus_1 = age + 1;
  los_plus_1 = LOS + 1;
  log_los= log(los_plus_1);
  log_totchg = log(totchg) ;
  log_nchronic = log(nchronic+1) ;
  *  Create indicators for primary payer ;
  if pay1 > . then do;
    Medicare = 0; Medicaid = 0; Private = 0; SelfPay = 0; NoCharge = 0;
OtherPay = 0;
    if pay1 = 1 then Medicare = 1 ;
      else if pay1 = 2 then Medicaid = 1;
      else if pay1 = 3 then Private = 1;
      else if pay1 = 4 then SelfPay= 1;
      else if pay1 = 5 then NoCharge = 1;
      else if pay1 = 6 then OtherPay = 1;
  end;
  *  Create indicators for race ;
  if race > . then do;
    white = 0; black = 0; hispanic = 0; asian = 0; native = 0; otherrace = 0;
    if race = 1 then white = 1;
    else if race = 2 then black = 1;
    else if race = 3 then hispanic = 1;
    else if race = 4 then asian = 1;
    else if race = 5 then native = 1;
    else if race = 6 then otherrace = 1;
  end;
  keep age age_plus_1 died hosp_nis hosp_division nis_stratum discwt orproc
female pay1 race los los_plus_1 log_los totchg log_totchg nchronic
log_nchronic
      missing_totchg missing_race missing_los missing_age missing_pay1
missing_female medicare medicaid private selfpay nocharge otherpay
      white black hispanic asian native otherrace ;
run;

*  Create age splines and log_los ;
```

```
proc transreg data=diabetes design;
  model bspl(age / knots=18 65 exknots=0 90);
  output out=age_splines ;
run;

*  Merge age splines to data ;

data diabetes ;
  merge diabetes age_splines(keep=age_0 age_1 age_2 age_3 age_4 age_5);
run;

*  Examine distibutions of binary and categorical variables ;

proc freq data=diabetes ;
  table female died pay1 race / missing ;
  Title 'Distribution of Categorical Variables, Complicated Diabetes, NIS
2012';
run;

*  Examine distibutions of continuous and count variables ;

proc univariate data=diabetes ;
  var totchg los age nchronic ;
  histogram totchg los age nchronic ;
  Title 'Distribution of Continuous Variables, Complicated Diabetes, NIS
2012';
run;

*  Output the missing value pattern ;

proc mi data=diabetes nimpute=0;
 var nis_stratum nchronic orproc female los age pay1 totchg race ;
 format age los nchronic orproc female pay1 race 5.2 totchg comma8.0;
 title 'Missing Data Patterns, Complicated Diabetes, NIS 2012';
run;

* Plot log(charges) vs age ;

proc summary data=diabetes nway;
  class age;
  var log_totchg ;
  output out=plotage mean= ;
run;

proc gplot data=plotage ;
  plot log_totchg * age ;
  symbol i=sm50 ; * Produce scatterplot smoother ;
 title 'Log(Charges) vs. Age, Complicated Diabetes, NIS 2012';
run;

* Plot log(charges) vs log(los+1) ;

proc summary data=diabetes nway;
  class log_los;
  var log_totchg ;
  output out=plotlos mean= ;
run;
```

```
proc gplot data=plotlos ;
  plot log_totchg * log_los ;
  symbol i=sm50 ; * Produce scatterplot smoother ;
 title 'Log(Charges) vs. Log(LOS+1), Complicated Diabetes, NIS 2012';
run;

* Plot log(charges) vs log(Nchronic+1) ;

proc summary data=diabetes nway;
  class log_nchronic;
  var log_totchg ;
  output out=plotchronic mean= ;
run;

proc gplot data=plotchronic ;
  plot log_totchg * log_nchronic ;
  symbol i=sm50 ; * Produce scatterplot smoother ;
 title 'Log(Charges) vs. Log(Chronic+1), Complicated Diabetes, NIS 2012';
run;

*  Regress log(totchg) on predictors to examine residuals ;
*  All cases with missing values are automatically dropped ;

ods graphics on;
proc reg data=diabetes ;
  model log_totchg = age_0 age_1 age_2 age_3 age_4 log_los log_nchronic
female
      medicaid private selfpay nocharge otherpay
      black hispanic asian native otherrace  ;
  var age ;
  output out=preds p=predicted r=residual ;
  title 'Regression of Total Charges on Predictors, Listwise Deletion,
Complicated Diabetes, NIS 2012';
run;

*  FCS method for imputation ;
*  Create 20 imputations, use SEED argument so that results can be
reproducted ;

*  Start by using MCMC to impute LOS, AGE, and FEMALE ;

proc mi data=diabetes nimpute=20 seed=57836763 out=imputed_1
 round = 1                    1                          1
 min =       1                      0                        1
 max =      91                      1                        361  ;
 var  age               female            LOS_plus_1 ;
 transform log(LOS_plus_1) ;
 mcmc chain=multiple nbiter=1000 niter=100 initial=em prior=jeffreys ;
run;

*  Merge age splines to data and calcualte log(LOS+1) for the next round of
imputations ;

proc transreg data=imputed_1 design;
  model bspl(age / knots=18 65 exknots=0 90);
  output out=age_splines ;
```

```
run;

data imputed_1 ;
  merge imputed_1 age_splines(keep=age_0 age_1 age_2 age_3 age_4 age_5);
  log_los = log(LOS_plus_1) ;
  LOS = LOS_Plus_1 - 1;
run;

*  Compare imputed to non-imputed values ;

proc freq data=imputed_1 ;
  table female * missing_female  ;
  Title 'Compare Observed and Imputed Percentage of Females' ;
run;

proc univariate data=imputed_1 ;
  class missing_age ;
  var age ;
  histogram age ;
  inset n mean median ;
  label missing_age = 'Missing Age';
  Title 'Compare Observed and Imputed Age Distribution' ;
run;

proc univariate data=imputed_1 ;
  class missing_los ;
  var log_los ;
  histogram log_los ;
  inset n mean median ;
  label missing_los = 'Missing LOS' log_los='Log(LOS+1)';
  Title 'Compare Observed and Imputed Log(LOS+1) Distribution' ;
run;


*  Create age splines and log_los ;

*  Look at missing value patterns after imputing LOS, age, and female ;

proc mi data=imputed_1 nimpute=0 ;
  var female age log_los race pay1 totchg ;
run;

*  Impute white and charges, include nis_stratum as a predictor ;

proc mi data=imputed_1 nimpute=1 seed=19520201 out=imputed_2 ;
 by _imputation_ ;
 *  Declare class variables ;
 class pay1 white nis_stratum ;
 *  Transform total charges for regressions ;
 transform log(totchg) ;
 *  Imputation model for primary white ;
 fcs logistic(white = age_0-age_4 female pay1 nis_stratum) ;
 *  Imputation model for primary payer ;
 fcs discrim(pay1 = age / prior=proportional) ;
 *  Imputation model for total charges ;
 fcs reg(totchg = female white pay1 age_0-age_4 log_los log_nchronic orproc
nis_stratum) ;
```

```
 *  Declare analysis variables ;
 var nis_stratum log_los age age_0-age_4 female orproc
      log_nchronic white pay1 totchg ;
 Title 'Impute Missing Values for White and Total Charges Using FCS,
Complicated Diabetes, NIS 2012';
run;

*  Calculate log(total charges) for comparisons ;

data imputed_2 ;
  set imputed_2 ;
  log_totchg = log(totchg) ;
run;

*  Compare imputed values to non-imputed values ;

proc freq data=imputed_2 ;
  table white * missing_race  pay1 * missing_pay1 ;
  title 'Compare Distributions Between Imputed and Non-Imputed Values,
Complicated Diabetes, NIS 2012';
run;

proc freq data=diabetes ;
 table pay1*(race missing_race) ;
  title 'Investigate the Relationship Between Primary Payer and Race,
Complicated Diabetes, NIS 2012';
run;

proc univariate data=imputed_2 ;
  class missing_totchg ;
  var totchg log_totchg ;
  histogram log_totchg ;
  inset n mean median;
  label log_totchg = 'Log(Total Charges)' missing_totchg = 'Missing Charges';
  title 'Compare Distributions Between Imputed and Non-Imputed Values,
Complicated Diabetes, NIS 2012';
run;


*  Calculate means and standard errors for each imputation ;

proc surveymeans data=imputed_2;
  by _imputation_ ;
  strata nis_stratum ;
  cluster hosp_nis ;
  weight discwt ;
  var totchg ;
  ods output statistics = imputed_means;
run;

*  Obtain pooled estimates.  For this analysis, EDF is 1 less than the number
of cases. ;

proc mianalyze data=imputed_means edf=105605;
  modeleffects Mean ;
  stderr StdErr ;
```

```
  Title 'Pooled Estimates Based on FCS Imputations, Complicated Diabetes, NIS
2012';
run;

*  Compare to estimated mean based on original data (listwise deletion) ;

proc surveymeans data=diabetes nmiss mean stderr;
  strata nis_stratum ;
  cluster hosp_nis ;
  weight discwt ;
  var totchg ;
  Title 'Estimate Based on Original Data (listwise deletion), Complicated
Diabetes, NIS 2012';
run;


*  Run regressions on original data (listwise deletion) ;

ods output ParameterEstimates=parms_original ;
proc glimmix data=diabetes noclprint  ;
  class hosp_nis nis_stratum ;
  model totchg = female white age_0-age_4 medicaid private selfpay nocharge
otherpay
      log_los log_nchronic orproc nis_stratum
     / dist=lognormal solution ddfm=BW ;
  random intercept / sub=hosp_nis type=VC ;
  Title 'Original Regression Estimates Based on Data with Missing Values,
Complicated Diabetes, NIS 2012';
run;

*  Create payer indicators for imputed data ;

data imputed_2 ;
  set imputed_2 ;
  if pay1 > . then do;
    Medicare = 0; Medicaid = 0; Private = 0; SelfPay = 0; NoCharge = 0;
OtherPay = 0;
    if pay1 = 1 then Medicare = 1 ;
      else if pay1 = 2 then Medicaid = 1;
      else if pay1 = 3 then Private = 1;
      else if pay1 = 4 then SelfPay= 1;
      else if pay1 = 5 then NoCharge = 1;
      else if pay1 = 6 then OtherPay = 1;
  end;
run;

*  Run regressions on imputed data ;

ods output ParameterEstimates=parms_by_imputation covb=covb_by_imputation;
proc glimmix data=imputed_2 noclprint ;
  by _imputation_ ;
  class hosp_nis pay1 nis_stratum ;
  model totchg = female white age_0-age_4 medicaid private selfpay nocharge
otherpay
      log_los log_nchronic orproc nis_stratum
     / dist=lognormal solution covb ;
  random intercept / sub=hosp_nis ;
```

```
  Title 'Imputation-Specific Regression Estimates Based on FCS Imputations,
Complicated Diabetes, NIS 2012';
run;

*  Create Pooled Estimates for the regression coefficients and their standard
errors ;

ods output ParameterEstimates=parms_pooled ;
proc mianalyze parms=parms_by_imputation edf=105395
               covb(effectvar=rowcol)=covb_by_imputation;
   modeleffects Intercept female white age_0 age_1 age_2 age_3 age_4
      medicaid private selfpay nocharge otherpay
      log_los log_nchronic orproc nis_stratum ;
  Title 'Pooled Regression Estimates Based on FCS Imputations, Complicated
Diabetes, NIS 2012';
run;

*  Compare original estimates to imputation-based estimates ;

data parameter_estimates ;
  merge parms_pooled(where=(Parm^='nis_stratum'))
           parms_original(where=(Effect^='NIS_STRATUM')
                 rename=(Estimate=Original_Estimate stderr=Original_StdErr
                 tValue=Original_tValue Probt=Original_Probt) drop=DF) ;
run;


proc print data=parameter_estimates label ;
  id Parm ;
  var Original_Estimate Estimate
        Original_StdErr StdErr
        Original_Probt Probt  ;
  label Estimate = 'Pooled Coefficient'
           Original_Estimate = 'Original Coefficient'
           StdErr = 'Pooled Std Error'
           Original_StdErr = 'Original Std Error'
           Probt = 'Pooled p-value'
           Original_Probt = 'Original p-value';
  format Original_Estimate Estimate 8.3 Original_StdErr StdErr 8.4 ;
  title 'Pooled vs. Original Regression Coefficients, Complicated Diabetes,
NIS 2012';
run;


*  Plot the multiplicative effect of age ;

*  Obtain age spline data ;

proc sort data=imputed_2 out=age_data_splines(keep=age age_0 age_1 age_2
age_3 age_4) nodupkey ;
 by age ;
run;

*  Put coefficients into a single record ;

proc transpose data=parms_original(where=(Effect ^= 'NIS_STRATUM'))
out=coeffs_original suffix=_orig;
```

```
  id effect ;
  var estimate ;
run;

proc transpose data=parms_pooled(where=(parm ^= 'nis_stratum'))
out=coeffs_pooled suffix=_pool;
  id parm ;
  var estimate ;
run;

*  Calculate multiplicative effects ;

data plot_age_data ;
  set age_data_splines ;
  if _n_=1 then do;
      set coeffs_original(keep=age_0_orig age_1_orig age_2_orig age_3_orig
age_4_orig) ;
      set coeffs_pooled(keep=age_0_pool age_1_pool age_2_pool age_3_pool
age_4_pool) ;
  end;
  retain age_0_orig age_1_orig age_2_orig age_3_orig age_4_orig
             age_0_pool age_1_pool age_2_pool age_3_pool age_4_pool;
  age_effect_original = exp(age_0 * age_0_orig + age_1 * age_1_orig + age_2 *
  age_2_orig + age_3 * age_3_orig + age_4 * age_4_orig) ;
  age_effect_pooled = exp(age_0 * age_0_pool + age_1 * age_1_pool + age_2 *
  age_2_pool + age_3 * age_3_pool + age_4 * age_4_pool) ;
run;

*  Plot multiplicative effects ;

axis1 label=(angle=90 'Multiplicative Effect of Age on Total Charges') ;
legend1 label=('Regression') value=('Original' 'Pooled');
symbol1 i=join w=2 color=blue line=1;
symbol2 i=join w=2 color=red line=2;
proc gplot data=plot_age_data ;
  plot Age_effect_Original*Age Age_effect_Pooled*Age / overlay vaxis=axis1
legend=legend1 ;
  label age='Patient Age';
  Title 'Multiplicative Effect of Age, Original vs. Pooled Regression,
Complicated Diabetes, NIS 2012';
run;
```