Frequently Asked Questions
2009-2010 National Survey of CSHCN

MISSING DATA AND MULTIPLE IMPUTATION

What are missing data?

- Missing data are pieces of information that were not collected during the interview and are not available to analysts. Examples include "don't know" or "refused" responses to questions as well as "legitimate skips" (which occur when we never intended to collect that answer or we already knew the answer based on prior responses and thus did not need to ask the question) and "missing in error" (when the data are missing because of interviewer error, a system error that prevented the data from being recorded, or some other such reasons). Item nonresponse for variables like income is common in population surveys. Legitimate skips are supposed to be missing; the remainder of this FAQ concerns missing data other than legitimate skips.

Why are missing data a problem?

- For weighted population count estimates, missing data always result in an underestimate of population counts, and for rates, missing data can result in an underestimate if the missing cases are included in the denominator. For population estimates of percent distributions, missing data can present a problem if the variable being analyzed has a different distribution among those with missing data than among those with non-missing data. If there is some systematic reason for cases missing data—a reason that is related to the attribute being measured— then population estimates of percent distributions based on the non-missing data will be biased.

What is imputation?

- Imputation is a process by which an attempt is made to replace missing data with a plausible prediction of what the missing data value would have been, based on the data that were collected. Typically, a statistical model is developed to determine reliable predictors of the attribute that has missing data, which are then used to derive a predicted value of the attribute for those cases with missing data on the attribute.

What is multiple imputation?

- Multiple imputation is an imputation process that implicitly recognizes that imputed values contain a certain amount of uncertainty. Multiple imputation software generates a set of plausible values for the missing data point, and statistical analysis software such as SAS or SUDAAN incorporates the uncertainty of the imputed values by essentially inflating the standard errors of estimates to reflect the variability of the imputed values across the set of imputations.

MISSING DATA AND THE NATIONAL SURVEY OF CSHCN

What is missing in the National Survey of Children with Special Health Care Needs?

- There are missing data throughout the data set – most questions have a small number of "don't know" or "refused" responses. Most of the time, the percent of cases with missing data is very small, and the missing data can be ignored. For a few variables, missing data are a bigger problem that should not be ignored. Some variables typically have a larger-than-normal amount of missing data because of participants' reluctance to reveal the information (e.g., due to privacy concerns); other variables have more missing data because those questions appear at the end of a long interview, and some respondents terminate the interview before the end. In the 2009-2010 National Survey of Children with Special Health Care Needs (NS-CSHCN), the following variables experienced either or both of these issues: household income, household size, highest education in the household, primary language spoken in the household, and child's race and ethnicity.

What is the possible impact of missing data on analyses of the NS-CSHCN data?

- Because the NS-CSHCN first screens all children in the household for special health care needs, and then randomly selects one child with special needs to be the target of the detailed interview, the interview length is considerably longer for households that include CSHCN than it is for households without CSHCN. Because items at the end of the longer interview suffer greater item nonresponse due to interview breakoff, the result is that missingness on these items is greater among households with CSHCN than among households without CSHCN. For analysis of variables in the interview file itself (i.e., analysis among CSHCN only), this does not necessarily pose a problem. But for estimates of the prevalence of special health care needs or other analyses of the screener or household files that include both children with and without special health care needs, the magnitude of missing data will be nonrandom by special health care needs status. This can result in the nonsensical situation where, for example, an estimate of the prevalence of special health care needs can be higher for the total population than it is for each and every subgroup defined by an income category, because prevalence was highest among those with missing income data. Estimates of the prevalence of special health care needs should not be calculated for subgroups defined by the variables that exhibit appreciable differential missingness by special health care needs status.

Did differential missingness impact estimates from earlier iterations of the survey?

- The 2009-2010 iteration of the NS-CSHCN is the first time we have encountered this specific problem, because we moved the questions on household and child demographics to the end of the interview for this iteration. Earlier iterations of the survey did have missing data, but they did not have this particular problem of missingness on key demographic variables that was nonrandom with respect to special health care needs status.

What has been done to address this problem with missing data?

- SLAITS has made available for public download two data files that contain multiply-imputed data for the 2009-2010 NS-CSHCN. The imputed household file includes imputed data for household income relative to the poverty threshold, household size, highest education in the household, and primary language spoken in the household. The imputed screener file includes imputed data for child's race and ethnicity. Each file contains a flag variable identifying which cases have imputed data. Both files contain values for all cases in the data – those without missing data are shown with the original valid values, while those originally with missing data have had that missing data replaced with imputed data. The files contain a set of five imputations for each NS-CSHCN data record, or a total of (196,159)(5) = 980,795 records in the imputed household file and (371,617)(5) = 1,858,085 records in the imputed screener file. Analysts can merge these imputed files to the publicly-released Screener, Household, and Interview files to do analyses that remove the bias associated with differential missingness by special health care needs status. Statistical software products such as SAS and SUDAAN can be used to accomplish the analysis while accounting for the uncertainty in the imputation process.

USING SAS AND SUDAAN TO ANALYZE MULTIPLY IMPUTED DATA

To merge imputed child race/ethnicity from the imputed screener file and imputed household income (for example) from the imputed household file with data from the Interview file (for estimates of insurance coverage for CSHCN, for example), first merge child interview data with child screener data by unique child identifier (IDNUMXR), and then merge that merged file to the household data by unique household identifier (IDNUMR).

```
DATA iscreener; /*IMPUTED SCREENER FILE*/
 SET CSHCN09.cshcn0910_multimp_screener;
 KEEP idnumxr imputation racer_i hispanic_i;
RUN;
PROC SORT data = iscreener;
 by idnumxr;
RUN;
```

```
DATA interview; /*INTERVIEW FILE*/
 SET CSHCN09.puf_cshcn_interview_unformat;
 KEEP idnumxr idnumr state sample weight_i;
RUN;
PROC SORT data = interview;
 by idnumxr;
RUN;

DATA iscr_int; /*MERGED IMPUTED SCREENER AND INTERVIEW*/
 MERGE iscreener interview (in = i);
 by idnumxr;
 if i;
RUN;
PROC SORT data = iscr_int;
 by idnumr imputation;
RUN;

DATA ihousehold; /*IMPUTED HOUSEHOLD FILE*/
 SET CSHCN09.cshcn0910_multimp_household;
 KEEP idnumr imputation povlevel_i;
RUN;
PROC SORT data = ihousehold;
 by idnumr imputation;
RUN;

DATA ihh_iscr_int; /*MERGED IMPUTED HOUSEHOLD, IMPUTED SCREENER, INTERVIEW*/
 MERGE ihousehold iscr_int (in = i);
 by idnumr imputation;
 if i;
RUN;
```

Be sure to include any additional analytic variables of interest in the KEEP statements below. If you also need to use unimputed data from the Screener or Household files, first merge Screener and imputed screener by IDNUMXR before merging to Interview, and first merge Household and imputed household by IDNUMR before merging to the Screener/Interview file.

To merge imputed child race/ethnicity from the imputed screener file and imputed household income from the imputed household file with data from the Screener file (for estimates of prevalence of special health care needs, for example), merge Screener and imputed screener data, then merge Household and imputed household data, then merge together.

```
DATA iscreener; /*IMPUTED SCREENER FILE*/
 SET CSHCN09.cshcn0910_multimp_screener;
 KEEP idnumxr imputation racer_i hispanic_i;
RUN;
PROC SORT data = iscreener;
 by idnumxr;
RUN;

DATA screener; /*SCREENER FILE*/
 SET CSHCN09.puf_cshcn_screener_unformat;
 KEEP idnumxr idnumr state sample weight_s needtype;
RUN;
PROC SORT data = screener;
 by idnumxr;
RUN;

DATA iscr_scr; /*MERGED IMPUTED SCREENER AND SCREENER*/
 MERGE iscreener screener;
 by idnumxr;
RUN;
```

```
PROC SORT data = iscr_scr;
 by idnumr imputation;
RUN;

DATA ihousehold; /*IMPUTED HOUSEHOLD FILE*/
 SET CSHCN09.cshcn0910_multimp_household;
 KEEP idnumr imputation povlevel_i;
RUN;
PROC SORT data = ihousehold;
by idnumr;
RUN;

DATA household; /*HOUSEHOLD FILE*/
 SET CSHCN09.puf_cshcn_household_unformat;
 KEEP idnumr state sample weight_h;
RUN;
PROC SORT data = household;
 by idnumr;
RUN;

DATA ihh_hh; /*MERGED IMPUTED HOUSEHOLD AND HOUSEHOLD*/
 MERGE ihousehold household;
 by idnumr;
RUN;
PROC SORT data = ihh_hh;
 by idnumr imputation;
RUN;

DATA ihh_iscr_hh_scr; /*MERGED SCREENER AND HOUSEHOLD WITH IMPUTATIONS*/
 MERGE iscr_scr ihh_hh;
 by idnumr imputation;
RUN;
```

To analyze the merged data in SUDAAN, the data set must be first separated such that the five imputations appear in five data sets, each with the same file name except the final number 1 through 5, and each sorted by the SUDAAN NEST statement variables:

```
DATA mimp1 mimp2 mimp3 mimp4 mimp5; /*CREATE IMPUTED FILES FOR SUDAAN*/
 SET ihh_iscr_hh_scr;
 if imputation = 1 then output mimp1;
 if imputation = 2 then output mimp2;
 if imputation = 3 then output mimp3;
 if imputation = 4 then output mimp4;
 if imputation = 5 then output mimp5;
RUN;
PROC SORT data=mimp1; by state sample idnumr; RUN;
PROC SORT data=mimp2; by state sample idnumr; RUN;
PROC SORT data=mimp3; by state sample idnumr; RUN;
PROC SORT data=mimp4; by state sample idnumr; RUN;
PROC SORT data=mimp5; by state sample idnumr; RUN;
```

Then, to analyze these five data sets in SUDAAN, simply add "data=mimp1" and "mi_count=5" to your usual PROC statement as shown here:

```
PROC CROSSTAB data=mimp1 design=wr mi_count=5 filetype=sas;
 NEST        state sample idnumr / psulev=3;
 WEIGHT      weight_s;
 CLASS       racer_i povlevel_i needtype;
 TABLES      racer_i*needtype povlevel_i*needtype;
 PRINT       nsum wsum rowper serow lowrow uprow / style=nchs;
RUN;
```

FURTHER INFORMATION

Whom do I contact if I have questions about imputation of NS-CSHCN data after I read this document?

- We recognize that this summary may not provide all of the information that analysts need regarding the development and use of the multiply imputed data. If you have further questions, please send an email to **slaits@cdc.gov**.

What is the suggested citation for this document?

- Centers for Disease Control and Prevention, National Center for Health Statistics, State and Local Area Integrated Telephone Survey. 2009-2010 National Survey of CSHCN Frequently Asked Questions: Missing Data and Multiple Imputation. March 2012. Available from URL: **http://www.cdc.gov/nchs/slaits/cshcn.htm**