

October 2018

Variance Estimation Guidance, NHIS 2016-2017 (Adapted from NHIS Survey Description Documents)

Introduction

The data collected in the NHIS are obtained through a complex sample design that involves stratification and clustering. The final weights provided for analytic purposes have been adjusted in several ways to permit calculation of valid estimates for the civilian, noninstitutionalized population of the United States. As with any variance estimation methodology, the techniques presented here involve several simplifying assumptions about the design and weighting scheme applied to the data. The first part of this document provides guidelines for data users based on simplified concepts of the NHIS sample design structure so that users may compute reasonably accurate standard error estimates. The second part of this document provides guidance for merging NHIS data files, and combining NHIS data across years.

Several software packages are available for analyzing complex samples. The website *Summary of Survey Analysis Software*, currently located at:

<http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html> ,

provides references for and a comparison of different software alternatives for the analysis of complex data. Analysts at NCHS generally use the software package SUDAAN[®] (Research Triangle Institute 2012) to produce standard error estimates. In this document, examples of SUDAAN computer code for standard error calculation are provided for illustrative purposes. Examples also are provided for the Stata, SPSS, SAS, R, and VPLX software packages. NCHS recommends that NHIS data be analyzed under the direction of or in consultation with a statistician who is cognizant of sampling methodologies and techniques for the analysis of complex survey data.

Users are reminded that the use of standard statistical procedures that are based on the assumption that data are generated via simple random sampling (SRS) generally will produce incorrect estimates of variances and standard errors when used to analyze data from the NHIS. The clustering protocols that are used in the multistage selection of the NHIS sample require other analytic procedures, as described below. Analysts who apply SRS techniques to NHIS data generally will produce standard error estimates that are, on average, too small, and are likely to produce results that are subject to excessive Type I error.

Conceptual NHIS design for 2017

The 2017 NHIS sample came from the second year of a new sample design that was implemented in 2016. This document provides a brief outline of the new NHIS sample design.

To achieve sampling efficiency and to keep survey operations manageable, cost-effective, and timely, the NHIS survey planners used complex sampling techniques to select the sample of dwelling units for the NHIS. These methods partition the target universe into strata and clusters.

The NHIS target universe is defined as all dwelling units in the U.S. that contain members of the civilian noninstitutionalized population (households and noninstitutional group quarters such as college dormitories). As the NHIS is conducted in a face-to-face interview format, a simple random sample of dwelling units would be too dispersed throughout the nation; as a result, the costs of interviewing a simple random sample of 35,000 dwelling units would be prohibitive. To achieve survey objectives subject to resource constraints, the NHIS uses methods of clustering and stratification.

First, the target universe was partitioned into groups of addresses located in geographic areas within each state, which are single counties, groups of adjacent counties (or equivalent jurisdictions), or metropolitan areas. The geographic areas were classified into one or two groups within each state. Within each state/geographic area group, a systematic sample of address clusters was selected.

The hierarchy of sampling allows the creation of household- and person-level base weights. Each address-level base weight is the inverse of the probability of selection of the address cluster containing the address. Roughly speaking, the base weight is the number of population units a sampled unit represents. Under ideal sampling conditions, and if 100% response occurred, a base-weighted sample total will be an unbiased estimator for the true total in the target population. In practice, however, the base weights are adjusted for non-response, and ratio-adjusted to create final sampling weights. The final person-level weights are adjusted according to a quarterly poststratification by age/sex/race/ethnicity classes based on population estimates produced by the U.S. Census Bureau that also are used for Current Population Survey weight adjustments. Most other weights receive some form of ratio adjustment as well.

Internally, NCHS uses the design and weighting information to formulate appropriate variance estimates for NHIS statistics. While recognizing the need to provide accurate information, NCHS also must adhere to the Public Health Service Act (Section 308(d)), which forbids the disclosure of any information that may compromise the confidentiality promised to its survey respondents. Consequently, much of the NHIS design information cannot be publicly released, and other data are either suppressed or recoded to insure confidentiality. In order to satisfy this disclosure constraint, many of the original design strata and clusters are masked for public release by applying techniques to cluster, collapse, mix, and partition the original design variables. Through this process the original NHIS design variables are transformed into public use design pseudo-variables (i.e., PSTRAT and PPSU). Data users who want access to internal NCHS data have the option of accessing data through the NCHS Research Data Center. For further information, refer to <http://www.cdc.gov/rdc>.

Design Information Available on the NHIS Public Use Data Files

The 2017 Household, Person, Sample Adult, Sample Child, and supplement public use files contain the design variables necessary for variance estimation. PSTRAT is the stratum for variance estimation variable; PPSU is the PSU for variance estimation variable; and WTFA is the weight variable. The stratum and PSU variable names are the same in the other files, but the weight variable has a different name.

As discussed above, in order to mask true geographical locations, the PSTRAT and PPSU

levels are pseudo-levels or simplified versions of the true NHIS sample design variables. Analysts are cautioned that these simplified design structures are not designed to support geographical analyses below the Census Region level.

The PSTRAT and PPSU values for 2017 are based on the current NHIS sample design, and have no connection with the STRAT_P and PSU_P values for 2006-2015. Refer to the section "Variance Estimation for Pooled Analyses of Adjacent Years of the NHIS" in this document for variance estimation guidance for pooled analyses of adjacent years of the NHIS, including pooling 2017 data with data for 2016 and earlier years.

Variance Estimation Method for Public Use Data

The method described below is applicable to the 2017 NHIS Household, Person, Sample Adult, and supplement files.

The limited public release design information requires a mathematical simplification that the PSUs be treated as if they were sampled with replacement (WR). This public use method tends to provide slightly more conservative (larger) standard errors than the variance estimation method that is applied internally by analysts at NCHS. The public use method is robust when analyzing subsetted or subgroup data (see the section "Subsetted Data Analysis" below).

The simplified design structure can be specified with the following statements in SUDAAN for the Person File:

```
PROC <DESCRIPT, CROSSTAB, ...>...DESIGN = WR ;  
NEST PSTRAT PPSU ;  
WEIGHT WTFA ;
```

Note that SUDAAN requires that the input file be sorted by the variables listed on the NEST statement (i.e., PSTRAT and PPSU). Design statements for other data files should use the appropriate weight variables found on these files.

Corresponding statements for other software packages are as follows:

Stata svy:

SVYSET [**PWEIGHT=WTFA**], *STRATA*(**PSTRAT**) *PSU*(**PPSU**)

SVY: MEAN <name of variable to be analyzed for average>

Or

SVY: PROPORTION <name of variable to be analyzed for percentage/proportion>

SPSS csdescriptives (for averages) or cstabulate (for percentages/proportions):

One needs first to define a "plan file" with information about the weight and variance estimation, e.g.:

CSPLAN ANALYSIS

/PLAN FILE="< file name >"

/PLANVARS ANALYSISWEIGHT=WTFA

/DESIGN STRATA=PSTRAT CLUSTER=PPSU

/ESTIMATOR TYPE=WR.

And then refer to the plan file when using *csdescriptives* or *cstabulate*, e.g.:

CSDESCRIPTIVES

/PLAN FILE="< file name >"

/SUMMARY VARIABLES =<name of variable to be analyzed>

/MEAN.

CSTABULATE

/PLAN FILE="< file name >"

/TABLES VARIABLES =<name of variable to be analyzed>

/CELLS TABLEPCT.

SAS proc surveymeans (for averages) or surveyfreq (for percentages/proportions):

PROC SURVEYMEANS;

STRATA PSTRAT;

CLUSTER PPSU;

WEIGHT WTFA;

VAR <name of variable to be analyzed>;

RUN;

PROC SURVEYFREQ;

STRATA PSTRAT;

CLUSTER PPSU;

WEIGHT WTFA;

TABLES <name of variable to be analyzed>;

RUN;

R (including the "survey" add-on package):

(note: R syntax is case-sensitive)

```
# load survey package
require(survey)
# create data frame with NHIS design information, using existing data frame of NHIS data
nhissvy <- svydesign(id=~ppsu, strata=~pstrat,
nest = TRUE,
weights=~wtfa,
data=< existing data frame name>)
svymean(~<name of variable to be analyzed>, design=nhissvy)
```

note: svymean will produce proportions for "factor variables". For details consult the R documentation (<http://cran.r-project.org/manuals.html>).

VPLX:

In the CREATE step, include the following statements:

```
STRATUM    PSTRAT
CLUSTER    PPSU
WEIGHT     WTFA
```

Then specify the variable to be analyzed in the DISPLAY step:

```
LIST      MEAN(<name of variable to be analyzed>)
```

VPLX can produce percentages by including a CAT statement in the CREATE step. For details consult the VPLX documentation (<http://www.census.gov/sdms/www/vdoc.html>).

A rule of thumb to calculate the number of degrees of freedom to associate with a standard error is the quantity *number of PSUs - number of strata*. Typically, this rule is applied to a design with two PSUs per stratum and when the variance components by stratum are roughly the same magnitude. The applicability of this rule of thumb depends upon the variable of interest and its interaction with the design structure (for additional information, see Chapter 5 of Korn and Graubard 1999). The number of degrees of freedom is used to determine the *t*-statistic, its associated percentage points, p-values, standard error, and confidence intervals. As the number of degrees of freedom becomes large, the distribution of the *t*-statistic approaches the standard normal distribution. For example, with 120 degrees of freedom, the 97.5 percentage point of the t_{120} distribution is 1.980, while the 97.5 percentage point of the standard normal distribution is 1.960. If a variable of interest is distributed across most of the NHIS sample areas, a normal distribution assumption may be adequate for analysis since the number of degrees of freedom would be large. The user should consult a mathematical statistician for further discussion.

Subsetted Data Analysis

Frequently, studies using NHIS data are restricted to specific population subgroups, e.g.,

persons aged 65 and older. Some users delete all records outside of the domain of interest (e.g., persons aged less than 65 years) in order to work with smaller data files and run computer jobs more quickly. This procedure of keeping only selected records (and list-wise deleting other records) is called subsetting the data. With a subsetted dataset that is appropriately weighted, correct point estimates (e.g., estimates of population subgroup means) can be produced. **However, in general, software packages that correctly analyze complex survey data cannot compute accurate standard errors for subsetted data.** When complex survey data are subsetted, oftentimes the sample design structure available to the software is incomplete; subsetting data deletes important design information needed for variance estimation. Note that SUDAAN has a SUBPOPN option that allows the targeting of a subpopulation while using the full (unsubsetted) data file containing the design information for the entire sample. (See a SUDAAN manual for more information.) **NCHS recommends that subpopulation analyses be carried out using the full data file and the SUBPOPN option in SUDAAN, or an equivalent procedure (see below) with another complex design variance estimation software package.**

Strategy 1 (recommended) Use the SUBPOPN statement with the method described above for the full Person File dataset:

```
PROC ...DESIGN = WR ;  
NEST PSTRAT PPSU ;  
WEIGHT WTFA ;  
SUBGROUP(variable names);  
LEVELS ...;  
SUBPOPN RACRECI3=2 & SEX=2 / NAME="Analysis of African American women;"
```

Using the full dataset with the SUBPOPN statement in this example would constrain this analysis to African American women only (RACRECI3 = 2 for black and SEX = 2 for female). Use of the SUBPOPN statement is equivalent to subsetting the dataset, except that any resulting variance estimates are based on the full design structure for the complete dataset.

Strategy 2 (not recommended, except when Strategy 1 is infeasible) Use the MISSUNIT option on the NEST statement with the method described above for subsetted data:

```
NEST PSTRAT PPSU / MISSUNIT ;
```

In a WR design, when some PSUs are removed from the database through the listwise deletion of records outside the population of interest, leaving only one PSU in one or more strata, the MISSUNIT option in SUDAAN "fixes" the estimation to avoid errors due to the presence of strata with only one PSU. In the special case of a WR design with exactly two PSUs per stratum, using the MISSUNIT option with subsetted data gives the same variance estimate as using Strategy 1. However, except for this special case, there is no guarantee that the variance estimates obtained by this method are equivalent to those obtained using Strategy 1. Other calculations, such as those for design effects, degrees of freedom, standardization, etc., may need to be carried out differently. Users are responsible for verifying the correctness of their results based on subsetted data.

Implementing Strategy 1 in other software packages can be accomplished as follows:

Stata svy:

Add SUBPOP to the SVY statement, e.g.:

```
SVY,SUBPOP( RACRECI3==2 & SEX==2 ): MEAN <name of variable to be analyzed>
```

SPSS csdescriptives or cstabulate:

One must first define an indicator variable, e.g.:

```
DO IF (RACRECI3 EQ 2 AND SEX EQ 2).  
  COMPUTE SUBGRP=1.  
ELSE.  
  COMPUTE SUBGRP=0.  
END IF.
```

And then refer to the indicator variable in csdescriptives or cstabulate, e.g.:

```
CSDESCRIPTIVES (or CSTABULATE)  
/SUBPOP TABLE=SUBGRP
```

It is **very important** that the indicator variable be defined for all data records, otherwise an invalid result can occur.

SAS proc surveymeans or surveyfreq:

One must first define an indicator variable, e.g.:

```
IF RACRECI3=2 & SEX=2 THEN SUBGRP=1;  
ELSE SUBGRP=0;
```

And then refer to the indicator variable in proc surveymeans using the DOMAIN statement, e.g.:

```
PROC SURVEYMEANS;  
DOMAIN SUBGRP;
```

Proc surveyfreq does not have a DOMAIN statement. Instead, include the indicator variable in the TABLES specification:

```
PROC SURVEYFREQ;  
TABLES SUBGRP*<name of variable to be analyzed>;
```

As with SPSS, it is **very important** that the indicator variable is defined for all data records; otherwise an invalid result can occur.

R (including the "survey" add-on package):

After applying the `svydesign` function to a data frame that contains the entire NHIS sample file being analyzed, create a new data frame using the criteria that define the subgroup of interest. Note that R is very "feisty" when testing for equality, hence the syntax that follows specifies the subgroup of interest without using an equality test.

```
# subset for racreci3=2 & sex=2 without using equal signs
subgrp <- subset(nhisvy,((racreci3>1) & (racreci3<3) & (sex>1)))
svymean(~<name of variable to be analyzed>,design=subgrp)
```

VPLX:

In the CREATE step, define one or more CLASS variables that can be used to specify the criteria that define the subgroup of interest.

```
COPY RACRECI3 INTO RACECAT
COPY SEX INTO SEXCAT
CLASS RACECAT (1/2/3-HIGH)
CLASS SEXCAT (1/2)
```

The second category of RACECAT, crossed with the second category of SEXCAT, defines the subgroup of interest.

Then, specify the variable to be analyzed in the DISPLAY step, and specify the subgroup of interest as well:

```
LIST      MEAN(<name of variable to be analyzed>) /CLASS RACECAT(2)*SEXCAT(2)
```

Note that the specification of RACECAT(2) and SEXCAT(2) is to the second category of each variable, which happens to be the value "2" in both cases in this example. Specification of RACECAT(3) would include all values of RACRECI3 of 3 and higher ("3-HIGH").

Variance Estimation for Pooled Analyses of Adjacent Years of the NHIS

Adjacent years of NHIS data sometimes are combined for a pooled analysis, e.g., 2015 and 2016, or 2012-2014. A pooled analysis might be done, for example, to increase the sample size for some small population. An estimate from a pooled analysis can be interpreted to be an estimate for the midpoint of or the "average" over the time interval of the pooled data.

See the section "Combining Years of Data" below in this document for an example SAS program that combines 2004 and 2005 NHIS data.

The sampling weights for pooled data should be adjusted; otherwise, estimates of totals will be too high. For example, the estimated total U.S. civilian noninstitutionalized population from two years of pooled data, using unadjusted weights, would be about twice as large as it should be. A simple, valid weight adjustment procedure that NCHS recommends is to divide each sample weight in the pooled dataset by the number of years that are being pooled; e.g., divide by 2 when two years of data are combined, divide by 3 when three years of data are combined, etc. A

sophisticated user may want to consider an alternative weight adjustment method that would minimize the variance of a particular estimate; however, in general, if the sample sizes are similar in the data years being combined, the simple procedure and the sophisticated alternative would give a similar adjustment.

Variance estimation for pooled analyses falls into one or more of the following three classifications:

- #1. The years being pooled fall within the same sample design period with the same public use design variables, and no changes were made to the design variables within the years being pooled.
- #2. The years being pooled fall into different sample design periods (e.g., design periods 1963-1972, 1973-1984, 1985-1994, 1995-2005, 2006-2015, 2016-2017).
- #3. The years being pooled fall within the same sample design period, and there were changes to the public use design variables (e.g., from 1995-1996 to 1997-2005).

For #1, the sample has been drawn from the same geographic areas (same sample design), and the definitions of the variables used for public use variance estimation have not changed within the time period being analyzed. A valid method for variance estimation is to treat the pooled data like one year of data with a very large sample size. It is not correct to treat the different data years as being statistically independent, because the samples for the different years were drawn from the same geographic areas. Treating different data years as being statistically independent generally will lead to standard error estimates that are too small, and standard error estimates of contrasts (differences) between years would tend to be too large if the yearly estimates are positively correlated.

For #2, the different sample design periods should be treated as statistically independent. If there are multiple years of data being used for one or both design periods, each group should be treated in a similar manner as described in #1, assuming that the design variables within each group were unchanged. For example, if 1992-1995 NHIS data were pooled, the #1 procedure applies for the 1992-1994 data, and that aggregate is treated as being statistically independent from the 1995 data.

Note that it may be necessary to create new design variables to carry out this type of analysis. For example, consider an analysis of 1992-1995 NHIS data. The design variables have different names in the two sample design periods, and the stratum identifiers have different lengths. Referring to the first method described in "Variance Estimation for Person Data Using SUDAAN and the National Health Interview Survey (NHIS) Public-Use Person Data Files, 1985-94", currently available online at <http://www.cdc.gov/nchs/nhis/sudaan.htm>, the (Method 1) design variables for the 1992-1994 data are CSTRATUM (stratum), CPSU (PSU), and WTF (weight), while they are STRATUM, PSU, and WFTA, respectively, for the 1995 data. Suppose the names of the new design variables are NSTRATUM (stratum), NPSU (PSU), and NWT (weight). One method to create values for NSTRATUM that are of consistent length and take account of the different sample design periods is to do the following: for the 1992-1994 data, where the CSTRATUM values are 1, 2, ..., 62, first change these to 001, 002, ..., 062 (consistent length with STRATUM), and then do something to make them distinct from the STRATUM values, such as put a "1" in front: 1001, 1002, ..., 1062. For the 1995 data, where the STRATUM values are 1, 2, ..., 339, first change these to 001, 002, ..., 339, and then do something to make them distinct from

the CSTRATUM values, such as put a "2" in front: 2001, 2002, ..., 2339. NPSU can be set equal to CPSU for the 1992-1994 data, and equal to PSU for the 1995 data, as both CPSU and PSU are of length one. NWT can be set equal to WTF/4 for the 1992-1994 data, and to WFTA/4 for the 1995 data.

In general, for classification #2, it is always necessary to create a new stratum variable, with distinct values for different sample periods. Adding multiples of 1000 to stratum variables within each sample design period is sufficient to assure distinct values. For example, if 2015 and 2016 data are combined, adding 1000 to the PSTRAT values in the 2016 data and then combining these modified values with the STRAT_P values in the 2015 data into a new stratum variable NSTRATUM is sufficient.

For #3, no entirely satisfactory approach is available. Grouping of years should be done over the periods where the same public use design variables are present (i.e., like #1). Then, for combining across years where there were changes to the public use design variables, the only option is to carry out an analysis as if the data years were statistically independent. For example, if 1995-1999 NHIS data were pooled, the #1 procedure applies for 1995-1996, and 1997-1999; then, the only alternative is to treat these two groups as statistically independent. The resulting standard error estimates may be too small, and standard error estimates of contrasts between years might be too large if the yearly estimates are positively correlated.

References

Cochran, W.G. (1977), *Sampling techniques* (3rd edition), John Wiley & Sons.

Korn, E.L., and Graubard, B.I. (1999), *Analysis of Health Surveys*, John Wiley & Sons.

National Center for Health Statistics (1999), *National Health Interview Survey: Research for the 1995-2004 redesign*, Vital and Health Statistics, Series 2, No. 126.

National Center for Health Statistics (2000), *Design and Estimation for the National Health Interview Survey, 1995-2004*, Vital and Health Statistics, Series 2, No. 130.

National Center for Health Statistics (2014), *Design and Estimation for the National Health Interview Survey, 2006-2015*, Vital and Health Statistics, Series 2, No. 165.

Research Triangle Institute (2012), *SUDAAN Language Manual, Volumes 1 and 2, Release 11*, Research Triangle Institute, Research Triangle Park, NC.

Merging Data Files and Combining Years of Data in the NHIS

NHIS data files can be merged within years as well as combined across years. The purpose of merging data *within* a particular data year is to incorporate variables from different data files when respondents are common to both files, thereby increasing the number of variables available for analysis for a given individual. In contrast, the purpose behind combining NHIS data files *across* survey years is to combine respondents from different data years while retaining variables common to both files, thereby increasing the number of respondents (as long as the same variables are found in both files) and the precision of estimates.

Merging Data Files

Once NHIS data files are sorted by record identifiers common to each file, merging is straightforward. Below is an example of a SAS program that will merge data files within an NHIS data year. Using the household, family, and person record identifiers (HHX, FMX and FPX, respectively), this program merges data from the 2012 Household, Family, Person, and Sample Child data files. **Variable names may change from one year to another. Users are advised to check variable names and update computer programs when changes occur.**

```
/* Merge the 2012 Household File and the 2012 Family File. */
```

```
/* Create a Household dataset with selected variables and sorted by HHX.*/
```

```
DATA HH (KEEP=HHX REGION); /* HH is a SAS dataset; the KEEP statement retains only  
the listed variables for processing. */
```

```
SET NHIS2012.HOUSEHLD; /*The SET statement reads data from the 2012 Household File.  
*/
```

```
PROC SORT DATA=HH; /* Sort by HHX, the household identifier. */
```

```
BY HHX;
```

```
RUN;
```

```
/* Create a Family dataset with selected variables and sorted by HHX. */
```

```
DATA FM (KEEP=HHX FMX INCGRP2 RAT_CAT2 WTFA_FAM); /* FM is a SAS  
dataset; the KEEP statement retains only the listed variables for processing. */
```

```
SET NHIS2012.FAMILYXX; /*The SET statement reads data from the 2012 Family File. */
```

```
PROC SORT DATA=FM; /* Sort by HHX, the household identifier. */
```

```
BY HHX;
```

```
RUN;
```

```
DATA HHFM; /* New combined dataset called HHFM */
```

```
MERGE FM (IN=FROMFM) HH; /* Merge the newly created FM and HH Files, using an IN  
statement.*/
```

```
BY HHX;
```

```
IF FROMFM = 1; /* The combined dataset HHFM will contain only those records that are in the  
Family File; the Household File's REGION variable will be appended to these records. */
```

```
PROC SORT DATA=HHFM; /* Sort by HHX and FMX, the household and family identifiers.  
*/
```

BY HHX FMX;
RUN;

In the code above, the IN statement creates a temporary SAS variable (called FROMFM) that has a value of 1 if the dataset associated with the IN statement contributed to the current observation or a value of 0 if it did not. The subsequent statement, “IF FROMFM = 1” tells SAS to retain only those observations from the Family File (called FM), thereby eliminating Household File records corresponding to non-response cases (no family/person records are available for non-response cases). For more information on IN statements in SAS, consult Delwiche and Slaughter (1998).

```
/* Merge the 2012 Person File and the combined 2012 Family/Household File. */
```

```
/* Create a Person File with selected variables. */
```

```
DATA PR (KEEP=HHX FMX FPX SEX AGE_P WTFA STRAT_P PSU_P); /* PR is a SAS  
dataset; the KEEP statement retains only the listed variables for processing. */  
SET NHIS2012.PERSONSX; /*The SET statement reads data from the 2012 Person File. */  
PROC SORT DATA=PR; /* Sort by HHX and FMX, the household and family identifiers. */  
BY HHX FMX;  
RUN;
```

```
DATA PRHHFM; /* Combined Person, Family, and Household dataset called PRHHFM*/  
MERGE PR HHFM (DROP=WTFA_FAM); /* Merge the newly created PR File and HHFM,  
the combined Family/Household File, by the identifiers common to both files. At this point, users  
may drop the Family File weight and retain only the Person File weight for person-level analyses.*/  
BY HHX FMX;  
PROC SORT DATA=PRHHFM; /* Sort by HHX, FMX, and FPX, the household, family, and  
person identifiers. */  
BY HHX FMX FPX;  
RUN;
```

The above code will create a person-level file, copying the family/household information to each matching person record.

```
/* Merge the 2012 Sample Child File and the combined 2012 Person/Family/Household File. */
```

```
/* Create a Sample Child File with selected variables. */
```

```
DATA CH (KEEP=FPX HHX FMX CASHMEV PROBRX WTFA_SC); /* CH is a SAS  
dataset; the KEEP statement retains only the listed variables for processing. */  
SET NHIS2012.SAMCHILD; /*The SET statement reads data from the 2012 Sample Child  
File. */  
PROC SORT DATA=CH; /* Sort by HHX, FMX, and FPX, the household, family, and person  
identifiers. */
```

```
BY HHX FMX FPX;  
RUN;
```

```
DATA CHPRHHFM; /* Combined Sample Child, Person, Family, and Household dataset called
```

CHPRHHFM*/

MERGE PRHHFM CH; /* Merge CH, the newly created Sample Child File, and PRHHFM, the combined Person/Family/Household File, by the identifiers common to both files. */

BY HHX FMX FPX;

RUN;

In the code above, no IN statement was used in the MERGE statement, so the resulting file will have records for all persons, sample child or not. The sample child data items will be missing for persons who do not have a matching sample child record.

Combining Years of Data

Important Note

Variable names may change from one year to another. Users are advised to check variable names and where names differ, make certain it is appropriate to combine years of data for a given variable.

As previously mentioned, the purpose of combining or concatenating years of data (in SAS terminology) is to increase the number of observations or respondents for the same number of variables and thus increase the precision of estimates. It is possible to combine data from successive years of the National Health Interview Survey (NHIS) when the questions remain essentially the same over the years being combined.

Combining datasets from more than one year joins them one after the other (concatenates), as opposed to merging datasets. Analysts wishing to do both – merge data from multiple files within years and combine years of data – will need to first merge the data within each single year and then concatenate the files for the selected years of data (see the preceding section on Merging Data Files).

Weights will normally need to be adjusted when combining data years. For example, if two years of NHIS data are combined, the sum of the weights will be about twice the size of the civilian noninstitutionalized population of the United States. To achieve annualized results when two years of NHIS data are combined, one method for weight adjustment is to divide each weight by two before analyzing the data.

If data from the period 2006-2015 are combined, the combined data should be treated like a single year of data with a larger sample size for the purpose of variance estimation. If data from the period 2006-2015 are combined with 2016-2017 data, variance estimation is more complicated, as data are being combined across two sample design periods. Refer to the section "Variance Estimation for Pooled Analyses of Adjacent Years of the NHIS" above for more information.

The following is an example of a SAS program that will combine data files across NHIS data years. The program is written to concatenate the data from the Person Files of the 2011 NHIS and the 2012 NHIS. This same program can be used to combine the 2004 and 2005 NHIS Person Files after minor modifications (e.g., change "2011" and "11" to "2004" and "04", change "2012" and "12" to "2005" and "05", and change STRAT_P PSU_P to STRATUM PSU).

Important Note

The person identifier was called PX in the 2003 (and earlier) NHIS and FPX in the 2004 (and later) NHIS; users may find it necessary to create an FPX variable in the 2003 and earlier datasets (or, alternatively, a PX variable in the 2004 and later datasets) in order to make the data compatible for analyses.

```
/*Combine data files from 2 different years. */
```

```
DATA PER_11; /* Create SAS dataset PER_11.*/  
SET NHIS2011.PERSONSX /* The SET statement reads data from an existing SAS dataset, e.g.,  
the 2011 Person File */ (KEEP=HHX FMX FPX AGE_P SEX WTFA STRAT_P PSU_P); /*  
The KEEP statement retains only the listed variables for processing. */  
RUN;
```

```
PROC SORT DATA=PER_11; /* Sort SAS dataset PER_11. */  
BY HHX FMX FPX;  
RUN;
```

```
DATA PER_12; /* Create SAS dataset PER_12.*/  
SET NHIS2012.PERSONSX /* The SET statement reads data from an existing SAS dataset, e.g.,  
the 2012 Person File */ (KEEP=HHX FMX FPX AGE_P SEX WTFA STRAT_P PSU_P); /*  
The KEEP statement retains only the listed variables for processing. */  
RUN;
```

```
PROC SORT DATA=PER_12; /* Sort SAS dataset PER_12. */  
BY HHX FMX FPX;  
RUN;
```

```
DATA COMB1112; /* New, combined SAS dataset */  
SET PER_11 PER_12; /* Concatenate selected variables from 2011 and 2012 datasets. */  
WTFA_2YR=WTFA/2; /*Create a new weight by dividing the existing Person File weight  
(WTFA) by 2, the number of Person data files combined to create the data file called COMBO.*/  
RUN;
```

Now, suppose there exists a dataset "COMB0405" with the combined 2004 and 2005 Person Files, and there exists a dataset "COMB1112" with the combined 2011 and 2012 Person Files. As part of creating a dataset named "COMB0412" containing the combined 2004, 2005, 2011 and 2012 Person Files, two issues need to be addressed:

1. Adjustment of weights
2. Formation of new variance estimation variables, because this combination goes across sample design periods

The weights in COMB0405 and COMB1112 should be divided by 2, so that the original weights have been divided by 4 (four years of data being combined). To avoid the possibility of errors, NCHS recommends that new names be used for the new variance estimation variables, e.g., NSTRATUM (stratum), NPSU (PSU). The PSU and PSU_P values from COMB0405 and COMB1112 can be copied directly to NPSU. The NSTRATUM values need to be created in such a way to assure the values are distinct between 2004-2005 and 2011-2012. As STRATUM ranges from 1 to 339 and STRAT_P ranges from 1 to 300, an appropriate method for creating the NSTRATUM values would be to add 1000 to the STRATUM values and 2000 to the STRAT_P values.

*/*Combine 2004-2005 data file with 2011-2012 data file */*

```
DATA COMB0405;  
SET COMB0405;  
DROP STRATUM PSU;  
NSTRATUM=STRATUM+1000;  
NPSU=PSU;  
RUN;
```

```
DATA COMB1112;  
SET COMB1112;  
DROP STRAT_P PSU_P;  
NSTRATUM=STRAT_P+2000;  
NPSU=PSU_P;  
RUN;
```

```
DATA COMB0412;  
SET COMB0405 COMB1112;  
DROP WTFA_2YR;  
WTFA_4YR=WTFA_2YR/2;  
RUN;
```

References

Delwiche, LD and SJ Slaughter. *The Little SAS Book: A Primer* (2nd edition), SAS Institute: Cary, NC. 1998.