Accounting for Sample Design in NHATS and NSOC Analyses: Frequently Asked Questions

Vicki A. Freedman
Mengyao Hu
Jill DeMatteis
Judith Kasper

Updated June 8, 2022

Suggested citation: Freedman, Vicki A., Mengyao Hu, Jill DeMatteis, Judith D. Kasper. 2022. Accounting for Sample Design in NHATS and NSOC Analyses: Frequently Asked Questions. NHATS Technical Paper #23 v2. Johns Hopkins University School of Public Health. Available at www.NHATS.org. This technical paper was prepared with funding from the National Institute on Aging (U01AG032947; R01AG054004).

Table of Contents

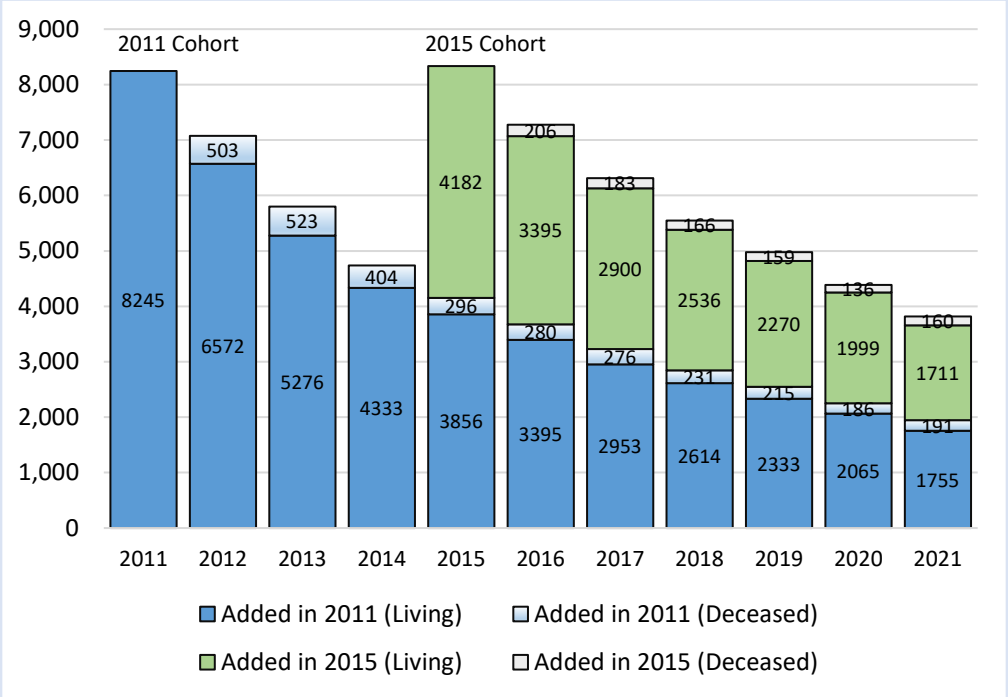
Introduction

The National Health and Aging Trends Study (NHATS) is designed to be nationally representative of persons 65 and older in the U.S. Started in 2011 and continuing with annual data collection, NHATS can be used to assess population-level trends (using repeat cross-sections) and to assess changes within individuals (following one or more cohorts over time).

The NHATS sampling design has several features that are characteristic of studies constructed to be nationally representative:

- NHATS is drawn from a sampling frame that includes the universe of persons of interest. The target populations for NHATS are cohorts of Medicare beneficiaries ages 65 and older in selected years (e.g. 2011, 2015, 2021, etc.).
- For feasibility and efficiency reasons, NHATS is geographically clustered. Counties are sampled from strata to ensure they are from each region of the country and that there is variation across counties in demographic composition. Within counties, zip codes (or clusters of zip codes) are then sampled.
- Because subgroup estimates by age and race are of interest, within these geographic clusters, persons at older ages and black individuals are oversampled relative to their representation on the Medicare frame.
- After the initial cohort (2011), each new cohort is formed by adding sample members to existing sample in such a way that preserves the ability to produce reliable subgroup estimates. That is, sample is added to all age/race groups. Consequently, cohort membership is overlapping (e.g. some members of the 2011 cohort are also members of the 2015 cohort, see the Figure 1 below).

Figure 1. NHATS Sample by Round



These sample design features—drawing from a frame, geographic clusters drawn from strata, over-sampling resulting in unequal probabilities of selection, and overlapping cohorts—have implications for how analytic samples are constructed and how estimation with common statistical software packages is specified. Analyses that draw upon repeated observations per individual may require additional considerations with regard to sample construction and estimation specification.

The National Study of Caregiving (NSOC) samples up to five caregivers among eligible NHATS sample persons (SP). This approach to generating the NSOC sample means that, in addition to the features above, caregivers in the sample are not independent observations but are clustered within NHATS SPs. This feature has additional implications for estimation.

For all analyses, researchers should be clear about the population to which they are making inferences (that is, to which cohort they are generalizing), apply the appropriate sampling weight (to take into account differential probabilities of selection and nonresponse) and specify the appropriate design variables (that reflect how are the geographic sampling units were drawn).

The rest of this document provides guidance on how to implement cross-sectional and longitudinal analyses that account for the complex sample design of NHATS and NSOC. Section I provides more detail on the basic design features of NHATS and NSOC. Section II provides an overview of weights. Section III offers guidance on cross-sectional analyses with NHATS and NSOC and Section IV focuses on longitudinal analyses with NHATS, including both trend analyses and within-individual changes. Sections III and IV offer examples using Stata (v. 15) and SAS (v. 9.4).

Section I: NHATS and NSOC Design Basics

What is the source of the NHATS sample?

NHATS uses the Medicare enrollment file as the sampling frame for selecting the NHATS sample. The file used to draw the sample was restricted to persons ages 65 and older enrolled in a given year (e.g. as of September 30, 2010 for the 2011 cohort, September 30, 2014 for the added sample in 2015 and so on).

How does the NHATS frame differ from the population 65 and older?

About 96%-97% of the population ages 65 and older in the US is enrolled in Medicare, making the enrollment file a very good approximation of the population ages 65 and older in the US.

Persons ages 65 and older not enrolled in Medicare include those who opt to stay on private insurance for some period after age 65 and delay enrolling and those who are ineligible (e.g. immigrated to the US after age 65). For more details see https://www.hhs.gov/answers/medicare-and-medicaid/who-is-elibible-for-medicare/index.html .

For cost reasons, before sampling, the frame was further limited to those living in the 48 contiguous states (excludes enrollees in AK, HI and US territories such as PR).

How is the NHATS sample clustered and how does clustering affect estimates?

The NHATS sampling design uses geographic clustering (selected across a number of strata) in order to make in-person data collection feasible on a national scale. To accomplish this step, the sampling frame was first clustered into Primary Sampling Units (PSUs) based on county. PSUs were then sampled from strata that were created based on region of the country and variables reflecting demographic composition. Within sampled PSUs, zip codes (or clusters of zip codes) were then sampled.

Clustering produces higher sampling errors than an unclustered sample of the same size because people in clusters (in this case within zipcodes) are more similar to one another and thereby have less variability on characteristics of interest than the population as a whole.

How do probabilities of selecting sample members differ in NHATS?

In NHATS probabilities of being selected from the frame into the sample differ by age and race. The sample is selected with the goal of roughly equivalent numbers across age groups: 65 to 69; 70 to 74; 75 to 79; 80 to 84; 85 to 89; 90+, which results in those at older ages being more likely to be selected, relative to their actual representation in the Medicare population.

In addition, Black non-Hispanic individuals (identified as such on the Medicare enrollment file) have a higher probability of selection relative to their representation. This differential sampling increases available sample for analyses that compare Black non-Hispanic and other individuals.

What cohorts are represented in NHATS?

NHATS samples overlapping cohorts of individuals ages 65 and older enrolled in Medicare. The 2011 cohort represents Medicare enrollees ages 65 and older as of September 30, 2010; the 2015 cohort represents enrollees ages 65 and older as of September 30, 2014; and the (planned) 2021 cohort will represent enrollees ages 65 and older as of September 30, 2020. Beginning in the second year of

interviewing for each cohort, deaths that have occurred since the last interview are captured each year in a separate Last Month of Life Interview.

Who is represented in each round (year) of NHATS?

Each round or year of NHATS is designed to be representative, when weighted, of the Medicare population at a given time, i.e., September 30 of 2010 for 2011 Cohort or September 30 of 2014 for 2015 Cohort , however the lower bound of the age range represented varies across rounds. In Round 1 (2011), the sample represents persons 65 and older, but the lower bound of the age range changes at each round with the aging of the sample, rising to 66 and older in Round 2, 67 and older in Round 3, and 68 and older in Round 4. In Round 5 (2015), when new sample was added (sample replenishment), the sample once again represents persons 65 and older (Round 6 represents those 66 and older and so on).

Does NHATS represent persons living in nursing homes and other residential care settings?

Yes, individuals are included in NHATS no matter where they reside. However, in the initial round of each cohort, if a person is determined to live in a nursing home [post-hospital Skilled Nursing Facility (SNF) stays are not considered evidence of nursing home residence], the sample person interview was not administered, only the facility questionnaire was completed (at the initial and all subsequent interviews). Individuals who move into nursing homes in subsequent years after participating in an SP interview in their initial round continue to be eligible for an SP interview.

Individuals who live in a nursing home when initially sampled can be identified using r#status on the Tracker file or r#dresid on the SP file, where # is the year they entered the sample (yearsample on the Tracker file).

Can NHATS be used to make national estimates of the number of older adults?

Yes, NHATS conducts interviews with a nationally representative sample of Medicare beneficiaries ages 65 or older, and therefore can be used to make national estimates of the number of older adults enrolled in Medicare as of the date the sample was drawn. However, because the sample was drawn in October preceding fieldwork (for details see Montaquila et al. 2012; DeMatteis et al. 2016), interviews do not begin until May of the following year and continue through early November, there is a gap of between 7 and 13 months between sampling and when individuals are interviewed. During this gap, the sample ages and deaths occur. Consequently, the resulting sample that is interviewed represents a slightly smaller and older population than the Medicare frame.

For many estimates of distributions within the older population and for the study of relationships among factors related to disability, these issues are not of consequence and can be ignored. Analysts interested in producing national estimates of the number of older adults with a particular characteristic may, however, wish to standardize their findings by age and sex to either the Medicare frame or, in some cases, to Census Bureau estimates. See NHATS technical paper #17 titled *"Making National Estimates with the National Health and Aging Trends Study"* (Freedman et al. 2016) for details on how to standardize NHATS estimates to the Medicare frame and Census totals.

What is the source of the NSOC sample?

NSOC has been conducted periodically in conjunction with NHATS. NSOC samples family and unpaid caregivers mentioned during the NHATS SP interview. If the SP received help with self-care, mobility or

household activities, the latter for health/functioning reasons, or if they lived in a residential care facility, and they have one or more family or unpaid caregivers, then all of the family and unpaid caregivers who provided help with activities are eligible. If more than 5 caregivers for a given SP are eligible, 5 caregivers are randomly sampled.

How do probabilities of selecting sample members differ in NSOC?

NSOC sample members have different probabilities of being selected that carry forward from NHATS (e.g. by the NHATS SP's age and race). In addition, those in networks larger than 5 have probabilities that decrease with network size.

How is the NSOC sample clustered?

Because the NSOC sample is drawn from NHATS, it is geographically clustered. In addition, NSOC caregivers may be clustered within NHATS SPs (if multiple caregivers for an NHATS SP respond).

Who is represented in each round of NSOC?

Each round of NSOC is designed to represent family and unpaid caregivers to the older Medicare population at a given time (when weighted). Cross-sectional files represent caregivers for persons with ages corresponding to the NHATS age profile in the round conducted and longitudinal files represent outcomes for a cohort of caregivers, beginning in Round 5 (2015) or Round 11 (2021).

Section II: NHATS and NSOC Sample Weights and Design Variables

Why are weights needed?

Weights have been developed each round to adjust for differential probabilities of selection and differential nonresponse. Persons with higher selection probabilities and higher chances of responding are assigned lower weights. Details of the procedures for developing weights for NHATS are provided in technical papers for each round; NSOC weights are described in the NSOC User Guide.

When weights are applied, the sample in a given round is representative of (and estimates generalizable to) the intended target population for the round. Without weights, the sample has too many SPs at older ages and black non-Hispanic individuals relative to their proportion in the population. In addition, nonresponse each round may result in biased estimates.

Some analysts prefer to follow the convention of controlling for variables used in the weights rather than weighting. In the case of NHATS the information that was used for oversampling (age and race on the Medicare enrollment sampling frame) and nonresponse adjustment (much of which is at the tract/county level) is not available on the public use files, and differs from round to round, so we do not encourage this approach.

What types of weights are available for NHATS at each round?

Three types of weights have been produced for each cohort each round:

- a ***base weight*** (on the Tracker file with the variable name w#trbaswgt*);
- a ***tracker weight*** (on the Tracker file with the variable name w#trfinwgt*); and
- an ***analytic weight*** (on the Sample Person file with the variable name w#anfinwgt* or w#an+wgt*)

where # is round, + is cohort (e.g. 2011 or 2015) and * is either 0 (full sample weight) or 1-56 (replicate weights).

The base weight accounts for differential probabilities of selection only. The tracker and analytic weights account for differential probabilities of selection and nonresponse. Most often analysts will be interested in using analytic weights applied to SP interviews which adjust for missing SP interviews for those in residential care settings.

When is it appropriate to use each type of weight?

Base weights that reflect only probabilities of selection in the year sampled are provided on the Tracker file. Base weights may be helpful for analysts who wish to explore their own nonresponse adjustments.

The tracker weights provided on the Tracker file are appropriate for making national estimates using the Facility Questionnaire (FQ) information (e.g. for services available to older adults living in residential care settings) since all persons with a completed FQ have a positive weight.

Most often analysts will be interested in using analytic weights found on the SP file. The analytic weights are appropriate for making national estimates using information obtained during SP interviews.

What is the difference between the tracker weights and the analytic weights?

There are two main differences between the tracker and analytic weights. Tracker weights are assigned to all cases with an SP interview (or a last month of life interview). Tracker weights are also assigned to cases in residential care who have an FQ interview only (missing an SP interview) and to cases who die between sample selection and the initial cohort interview. As a result, these weights are suited to particular analyses that do not require information from the SP interview (e.g. estimates of residential care generated from the FQ interview or mortality estimates generated from the tracker file).

Analytic weights are designed to be used for analyses of the SP interview, which include adjustments for the NHATS complex design and nonresponse, and also an additional nonresponse adjustment applied to cases in residential care with an SP interview (this accounts for cases in these settings that have only an FQ interview). All last month of life interviews also have an analytic weight.

Why is there more than one type of analytic weight on the SP file?

Analytic final weights (w#anfinwgt*, where # is round, and * is 0-56) are provided every round and always represent a specific round (i.e. year) of NHATS for the *most recent NHATS cohort* (e.g. 2011 or 2015). ***Analytic cohort weights*** (w#an+wgt*, where # is round, + is cohort, and * is 0-56) were created starting in 2015, to represent the population for *an earlier NHATS cohort*. Through 2020, there are two cohorts – 2011 and 2015 – starting in 2021, there will be two "earlier" cohorts – 2011 and 2015.

In other words, in the round where survivors of one cohort become members of a new cohort, these cases have weights for both cohorts for the current round: one that reflects the new cohort (w#anfinwgt*) and another that reflects membership in the earlier cohort (w#an+wgt*).

Whether analysts use the analytic final weights or the analytic cohort weights depends upon the population to which they are trying to generalize (e.g. survivors of the 2011 cohort, survivors of the 2015 cohort).

What is the difference between full sample weights and replicate weights?

For each set of NHATS weights, two types of weights have been produced: ***full sample weights*** and ***replicate weights***. The full sample weights have variable names that end in "0" and the replicate weights have variable names that end in values ranging from "1" to "56".

When Taylor series linearization estimation methods are used, the full sample weights should be specified (together with clustering and stratum variables, varunit and varstrat). When replication methods are used, both the replicate and full sample weights are specified. Since replicate weights reflect the geographic stratification and clustering, the clustering and stratum variables are not needed when using replicate weights.

What are the NHATS design variables that address geographic stratification and clustering and when are they needed?

When using the full sample weights with Taylor series linearization methods, analysts must also specify design variables that capture the effects on precision of estimates as a result of the geographic-based design. Two variables are included in the NHATS data for this purpose: one indicates Primary Sampling Unit (PSU) or cluster (varunit) and the other indicates the stratum from which the PSUs were drawn (varstrat).

Can multiple NHATS cohorts (e.g. 2015 and 2011) be combined? Can all observations from all Rounds of NHATS be combined?

For repeat cross sectional analyses (e.g. to estimate a trend) all rounds can be pooled using the analytic weight for each round.

For analyses that follow individuals over time (e.g. to estimate a trajectory or within individual changes), the specific cohort of interest should be specified before conducting the analysis. Weights that represent the cohort of interest should be used – i.e., the round-specific final weight for the initial years of the cohort and the cohort weight in rounds when a more recent cohort has begun.

For example, for analyses of the 2011 cohort in Rounds 1 through 4, analysts should use the analytic final weights (w#anfinwgt*, where # is round and * is 0-56). For analyses of the 2011 cohort in Round 5 and going forward, analyses should use the analytic cohort weights (w#an2011wgt*). See Section IV for further guidance on analyses of multiple rounds that take into account multiple observations per person.

What weights are available for NSOC at each round?

In NSOC I (2011), II (2015), and III (2017) cross sectional files with weights for caregivers to living SPs are provided. NSOC III also includes caregivers to deceased SPs and separate weights for these cases. NSOC III longitudinal files include both tracker and analytic weights for caregivers (who were also interviewed in 2015) to both living and deceased SPs.

For each weight, two types of weights have been produced: full sample weights and replicate weights. The full sample weights end in "0" and the replicate weights range from "1" to "56".

Both the full sample and replicate weights account for the NHATS design (including probabilities of selection and nonresponse), probabilities of selection for NSOC, and two types of nonresponse (SP's willingness to provide contact information and the caregiver's willingness to participate).

What are the NSOC design variables to address clustering?

When using the Taylor series linearization estimation method, analysts must also specify design variables that capture geographic stratification and clustering. Two variables are included in the NSOC data that reflect the geographic sampling that was implemented: Primary Sampling Unit (PSU) (c#varunit, where # is round) and sampling strata (c#varstrat, where # is round). Alternatively, analysts may wish to use the unique SP identifier "spid" as a cluster variable, since caregivers are clustered within NHATS SPs.

What weights are available for the NSOC III Time Diary interview?

NSOC III participants who reported providing help to a living NHATS Sample Person in the last month were offered a time diary follow-up interview. Summary information from the diary is available in a Summary File and information about each activity on the previous day is provided in the Activity File. The Summary File includes weights for the cross-sectional (t7diarywgt0-56) and longitudinal (lt7diarywgt0-56) samples that are designed to be used with diary-level analyses. The Activity File includes weights for the cross-sectional (t7dwbwgt0-56) and longitudinal (lt7dwbwgt0-56) samples that

are designed to support analysis of the random set of activities for which detailed wellbeing was collected. Both files include stratum and cluster variables (t7varstrat and t7varunit, respectively)

Section III. Single Round (Cross-sectional) Analyses

A single round of NHATS and/or NSOC may be of interest for some analyses. For example, analysts may wish to estimate the percentage of persons ages 65 and older receiving help with daily activities in 2015 or the percentage of family caregivers who helped with medications in 2017. For cross-sectional analyses that use a single round of data, both weight and design variables should be specified. Examples in this section assume that the Taylor series method of variance estimation is applied.

How do I specify the NHATS design when using the sample person interview?

Stata	svyset w#varunit [pweight=w#anfinwgt0], strata(w#varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight w#anfinwgt0; cluster w#varunit; strata w#varstrat; *[model or other statement]*; run;
R	library(survey) #need this line only once per session nhats.dsgn <- svydesign(id=~w#varunit, strata=~w#varstrat, weights=~w#anfinwgt0, data = *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	where #=current round number

How do I specify the NHATS design when using the last month of life interview?

Stata	svyset w#varunit [pweight=w#anfinwgt0], strata(w#varstrat) svy, subpop(fl#spdied==1): *[stata commands]*
SAS	*[sas procedure]*; weight w#anfinwgt0; cluster w#varunit; strata w#varstrat; *[model or other statement]*; domain fl#spdied; run;
R	library(survey) #need this line only once per session nhats.dsgn <- svydesign(id=~w#varunit, strata=~w#varstrat, weights=~w#anfinwgt0, data = *[data frame name]*, nest=TRUE) #users can use svyby() function for subpopulation analysis svyby(~[variable of interest], ~fl#spdied, nhats.dsgn, na.rm=T, [survey statistic to compute, e.g., svymean]) #users can also add subset() function within other commands for subpopulation analysis nhats.subset.dsgn <- subset(nhats.dsgn, fl#spdied == 1)
Notes	where #=current round number

How do I specify the NHATS design when using the facility interview?

Stata	svyset w#varunit [pweight=w#trfinwgt0], strata(w#varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight w#trfinwgt0; cluster w#varunit; strata w#varstrat; *[model or other statement]*; run;
R	library(survey) #need this line only once per session nhatsfc.dsgn <- svydesign(id=~w#varunit, strata=~w#varstrat, weights=~w#trfinwgt0, data = *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	where #=current round number

How do I specify the NSOC design when using cross-sectional NSOC files?

Analysts may wish to specify that caregivers are clustered geographically by SPs (using the varunit variable) or clustered within SPs (using spid). Alternatively if between SP variance is of interest, a multi-level model that controls for SP-level effects can be specified with caregivers as level 1 and SPs as level 2.

Note for NSOC I only, users should first recode c1varstrat=54 to c1varstrat=55 because c1varstrat=54 has no cases in cluster 1.

Stata	svyset c#varunit [pweight=w#cgfinwgt0], strata(c#varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight w#cgfinwgt0; cluster c#varunit; strata c#varstrat; *[model or other statement]*; run;
R	library(survey) #need this line only once per session nsoc.dsgn <- svydesign(id=~c#varunit, strata=~c#varstrat, weights=~ w#cgfinwgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	where #=current round number accounts for geographic clustering of caregivers

Stata	svyset spid [pweight=w#cgfinwgt0], strata(c#varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight w#cgfinwgt0; cluster spid; strata c#varstrat; *[model or other statement]*;

	run;
R	library(survey) #need this line only once per session nsoc.dsgn2 <- svydesign(id=~spid, strata=~c#varstrat, weights=~ w#cgfinwgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	where #=current round number accounts for clustering of caregivers within SPs

Stata	/*multilevel generalized linear model*/ /*specified as logit model*/ svyset c#varunit, weight(w#anfinwgt0) strata(c#varstrat) || _n, weight(cg_wt_r) svy: meglm *[outcome predictors]*, family(bernoulli) link(logit) || spid: /* specified as linear model*/ svyset c#varunit, weight(w#anfinwgt0) strata(c#varstrat) || _n, weight(cg_wt_r) svy: meglm *[outcome predictors]* || spid:
SAS	Not available
R	Not available
Notes	where #=current round number where cg_wt_r = w#cgfinwgt0 / w#anfinwgt0 multi-level models control for SP-level effects

How do I specify the NSOC design when using NSOC time diary files at the caregiver level?

The NSOC III time diary data are distributed in two files: a summary file and an activity file. The summary file is at caregiver level, containing all Round 7 NSOC cross-sectional and longitudinal time diary respondents. The activity file is at the activity level and includes more than 50,000 activities reported in more than 2,000 diaries.

For each file, analysts can use either the NSOC III cross-sectional or longitudinal samples or may pool the two samples for analysis. Analysts may also sum across activities to the caregiver level.

To use the time diary summary file, or information from the activity level file summed to the caregiver level, with cases in the NSOC III cross-sectional sample, you may either account for geographic clustering of the care recipients in NHATS or clustering of caregivers within NHATS recipients, as follows:

Stata	svyset t7varunit [pweight=t7diarywgt0], strata(t7varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight t7diarywgt0; cluster t7varunit; strata t7varstrat; *[model or other statement]*; run;
R	library(survey) #need this line only once per session nsoctd.dsgn <- svydesign(id=~t7varunit, strata=~t7varstrat, weights=~t7diarywgt0, data= *[data frame name]*, nest=TRUE)

	[model or other statement]
Notes	accounts for geographic clustering of caregivers

Stata	svyset spid [pweight=t7diarywgt0], strata(t7varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight t7diarywgt0; cluster spid; strata t7varstrat; *[model or other statement]*; run;
R	library(survey) #need this line only once per session nsoctd.dsgn2 <- svydesign(id=~spid, strata=~t7varstrat, weights=~t7diarywgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	accounts for clustering of caregivers within SPs

To use the time diary summary file or information from the activity level file summed to the caregiver level with cases in the NSOC III longitudinal sample, you may either account for geographic clustering of the care recipients in NHATS or clustering of caregivers within NHATS recipients, as follows:

Stata	svyset t7varunit [pweight=lt7diarywgt0], strata(t7varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]* weight lt7diarywgt0; cluster t7varunit; strata t7varstrat; *[model or other statement]*; run;
R	library(survey) #need this line only once per session lnsoctd.dsgn <- svydesign(id=~t7varunit, strata=~t7varstrat, weights=~t7diarywgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	accounts for geographic clustering of caregivers

Stata	svyset spid [pweight=lt7diarywgt0], strata(t7varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight lt7diarywgt0; cluster spid; strata t7varstrat; *[model or other statement]*; run;
R	library(survey) #need this line only once per session lnsoctd.dsgn2 <- svydesign(id=~spid, strata=~t7varstrat, weights=~t7diarywgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*

Notes	accounts for clustering of caregivers within SPs

To analyze the pooled sample consisting of both cross-sectional and longitudinal caregivers, you must first create a pooled weight. For cases in the cross-sectional sample only and cases in both cross-sectional and longitudinal samples, we recommend using the cross-sectional weight as the pooled weight. For cases that are only in the longitudinal file, we recommend using the longitudinal weight as the pooled weight.

Stata	svyset t7varunit [pweight=spl7diarywgt0], strata(t7varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]* weight spl7diarywgt0; cluster t7varunit; strata t7varstrat; *[model or other statement]*; run;
R	library(survey) #need this line only once per session pnsoctd.dsgn <- svydesign(id=~t7varunit, strata=~t7varstrat, weights=~spl7diarywgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	where spl7diarywgt0 is the created pooled weight; accounts for geographic clustering of caregivers

Stata	svyset spid [pweight=spl7diarywgt0], strata(t7varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight spl7diarywgt0; cluster spid; strata t7varstrat; *[model or other statement]*; run;
R	library(survey) #need this line only once per session pnsoctd.dsgn2 <- svydesign(id=~spid, strata=~t7varstrat, weights=~spl7diarywgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	where spl7diarywgt0 is the created pooled weight; accounts for clustering of caregivers within SPs

How do I specify the NSOC design when using the subset of activities in the NSOC time diary files with detailed wellbeing?

To analyze detailed wellbeing in the time diary activity file, analysts may use either the cross-sectional sample, the longitudinal sample, or the pooled sample (both cross-sectional and longitudinal cases). Here we provide an example of how to analyze the data using a pooled sample that takes into account clustering of activities within caregivers.

Analysts may specify the caregiver id (spid*100+opid) as the cluster variable, or if they are interested in variation in wellbeing across caregivers, they can adopt a multi-level modeling approach that

incorporates the NSOC time diary design. For the former, we recommend using the subpop command to subset the sample of activities to those with detailed wellbeing.

Stata	/*create unique id at caregiver level*/ egen cgid = group(spid opid), label svyset cgid [pweight=apl7diarywgt0], strata(t7varstrat) svy, subpop(if t7wdwb==1): *[stata procedures]*
SAS	/*create unique id at caregiver level*/ cgid = spid * 100 + opid; *[sas procedure]*; weight apl7diarywgt0; cluster cgid; strata t7varstrat; domain t7wdwb; *[model or other statement]*; run;
R	/*create unique id at caregiver level*/ *[data frame name]*$cgid <- paste0(*[data frame name]*$spid, *[data frame name]*$opid) library(survey) #need this line only once per session pnsoctd.wb.dsgn <- svydesign(id=~cgid, strata=~t7varstrat, weights=~apl7diarywgt0, data= *[data frame name]*, nest=TRUE) #users can add subset() function within other commands for subpopulation analysis pnsoctd.subwb.dsgn <- subset(pnsoctd.wb.dsgn, t7wdwb == 1)
Notes	where apl7diarywgt0 is the created pooled weight accounts for clustering of activities within caregivers

Stata	/*multilevel generalized linear model*/ /*specified as logit model*/ svyset t7varunit, weight(spl7diarywgt0) strata(t7varstrat) || _n, weight(plcgact_wt_r) svy, subpop(if t7wdwb==1): meglm *[outcome predictors]*, family(bernoulli) link(logit) || cgid: /* specified as linear model*/ svyset t7varunit, weight(spl7diarywgt0) strata(t7varstrat) || _n, weight(plcgact_wt_r) svy: meglm *[outcome predictors]* || cgid:
SAS	Not available
R	Not available
Notes	where plcgact_wt_r = apl7diarywgt0 / spl7diarywgt0, apl7diarywgt0 is the created pooled weight for activity file, and spl7diarywgt0 is the created pooled weight for summary file

	multi-level models control for caregiver-level effects

How do I specify the NHATS design for analyses that use propensity score weighting?

Propensity score weighting usually involves two stages: 1) estimating the propensity of receiving the treatment (that is, the "propensity score") and 2) estimating the treatment effects (that is, the effect of the predictor of interest on the outcome of interest), given the propensity score. Note that propensity score matching (as opposed to propensity score weighting) does not allow survey design to be easily incorporated.

When estimating the propensity score model it is appropriate to ignore survey design, since generalization to a population is not of interest at this stage. However, when estimating the treatment effect, it is important to take into account the NHATS survey design features, since generalization is of interest. The Stata code shown below is based on recommendations provided by Dugoff and colleagues[1], in which propensity score weights and survey weights are multiplied to form a new weight.

Stata	/*create propensity score, including survey weight as one of the covariates*/ *treatment is a binary variable indicating treatment group or control group pscore treatment [covariates] w#anfinwgt0, pscore(newvar1) blockid(newvar2) logit /*assess propensity score's balance across treatment and comparison groups*/ pbalchk treatment *[covariates]* /*weight treatment and comparison groups by the propensity score, using same covariates in previous two steps*/ qui dr outcome treatment *[covariates]* w#anfinwgt0, genvars *iptwt, the inverse probability of treatment weight will be generated egen *sumofweights* = total(iptwt) gen *norm_weights* = iptwt/sumofweights /*Multiply propensity score weight by survey weight*/ gen *newweight* = norm_weights*svyweight /*run outcome model*/ svyset w#varunit [pw=newweight], strata(w#varstrat) svy: *[Stata command]*
Notes	where # = current round number

[1] DuGoff, Eva H., Megan Schuler, and Elizabeth A. Stuart. 2014. Generalizing observational study results: applying propensity score methods to complex surveys. *Health Services Research* 49(1): 284-303.

Section IV. Analyses that Use Multiple Rounds

For analyses that involve more than one round of data, analysts must decide which weight to use and how best to specify design variables, including clustering. Which weight to use depends in part on the population to which the analyst wishes to generalize and whether loss to follow up (that is, non-response at future rounds) should be taken into account. Because many individuals appear in more than one round, analysts also may want to specify clustering within SP (instead of at the geographic level) and for some analyses may want to consider using a multi-level model. Examples in this section assume that Taylor series variance estimation method is applied.

How do I specify the NHATS design in analyses of trends over time for living SPs?

When analyzing repeated cross-sections (rounds) using NHATS, analysts should be aware of the following:

- Make sure to restrict the age range to persons at least age 70 or older since as the sample ages, representation of persons 65 to 69 decays (e.g. in Round 2 sample is 66 to 69; in Round 3 sample is 67 to 69).
- Make sure that at each round subjects are living in the same types of places (e.g. all rounds include or exclude all nursing home residents using r#dresid or r#status) and that all deceased cases are removed (using r#dresid or r#status).
- Because the same subjects appear across multiple rounds, specify that observations are clustered within SP (using spid).
- Use weights each round that take into account differential probabilities of selection and nonresponse.

Stata	svyset spid [pweight=wanfinwgt0], strata(w#varstrat) svy, subpop(keep_me==1): *[stata procedures]*
SAS	*[sas procedure]*; weight wanfinwgt0; cluster spid; strata w#varstrat; domain keep_me; *[model statements]*; run;
R	library(survey) #need this line only once per session nhats.trends.dsgn <- svydesign(id=~spid, strata=~w#varstrat, weights=~wanfinwgt0, data= *[data frame name]*, nest=TRUE) #users can add subset() function within other commands for subpopulation analysis nhats.subtr.dsgn <- subset(nhats.trends.dsgn, keep_me == 1)
Notes	where the data are pooled across rounds and wanfinwgt0 is set to w#anfinwgt0 and keep_me is set to 1 for observations to be kept in the analysis (e.g. by age, place and living/deceased status) where #=current round number accounts for clustering of observations within SPs

How do I specify the NHATS design in analyses of trends over time for deceased SPs?

When analyzing deaths over time using NHATS, several issues should be addressed:

- Make sure that you distinguish between the year the death is identified (round) and the year the death occurred (pd6yrdied on the sensitive file)
- Deaths found in replenishment years (e.g. Round 5, Round 12) represent continuing sample deaths of a prior cohort so be sure to use the appropriate cohort weight for replenishment years; in other years the final analytic weight for that round is appropriate.

Stata	svyset w#varunit [pweight=wanfinwgt0], strata(w#varstrat) svy, subpop(keep_me==1): *[stata procedures]*
SAS	*[sas procedure]*; weight wanfinwgt0; cluster w#varunit; strata w#varstrat; domain keep_me; *[model statements]*; run;
R	library(survey) #need this line only once per session nhats.lmltrends.dsgn <- svydesign(id=~w#varunit, strata=~w#varstrat, weights=~wanfinwgt0, data= *[data frame name]*, nest=TRUE) #users can add subset() function within other commands for subpopulation analysis nhats.sublmltr.dsgn <- subset(nhats.lmltrends.dsgn, keep_me == 1)
Notes	where the data are pooled across rounds wanfinwgt0 is set to w#anfinwgt0 for round 2 to 4, w5an2011wgt0 for round 5, and w#anfinwgt0 for round 6 to 8 keep_me is set to 1 for observations to be kept in the analysis by living/deceased status

How do I specify the NHATS design in analyses of individual change (trajectories) over time?

When analyzing repeated observations within subjects in an NHATS cohort, the same subjects appear in multiple rounds. We therefore recommend that analysts specify that observations are clustered within SP.

Analysts should first select a target population; for example they may select the 2011 cohort or the 2015 cohort, depending on the research question. They should then decide how to approach the issue of multiple (repeated) observations. They may either specify spid as the cluster variable or, if they are interested in what accounts for between-SP differences, they may choose to adopt a multi-level modeling approach that incorporates the NHATS sample design. For both approaches, the analyst should use weights each round that take into account differential probabilities of selection and nonresponse as well as cluster and strata variables.

Examples of 1) clustering on spid and 2) specifying a multi-level model are shown below:

Stata	svyset spid [pweight=wanfinwgt0], strata(w#varstrat)

	svy: *[stata procedures]*
SAS	*[sas procedure];* weight wanfinwgt0; cluster spid; strata w#varstrat; *[model statements];* run;
R	library(survey) #need this line only once per session nhats.traj.dsgn <- svydesign(id=~spid, strata=~w#varstrat, weights=~wanfinwgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	where the data are pooled across rounds for each cohort. For 2011 cohort, wanfinwgt0 is w#anfinwgt0 for round 1 to 4, and w#an2011wgt0 for round 5 to 9, where # = current round number. For 2015 cohort, wanfinwgt0 is w#anfinwgt0 for round 5 to 9. accounts for clustering of observations within SPs

Stata	/*multilevel generalized linear model for 2011 cohort*/ /*specified as logit model*/ svyset w#varunit, weight(w1anfinwgt0) strata(w#varstrat) || _n, weight(an2011_wt_r) svy: meglm *[outcome predictors]*, family(bernoulli) link(logit) || spid: /* specified as linear model*/ svyset w#varunit, weight(w1anfinwgt0) strata(w#varstrat) || _n, weight(an2011_wt_r) svy: meglm *[outcome predictors]* || spid: /*multilevel generalized linear model for 2015 cohort*/ /*specified as logit model*/ svyset w#varunit, weight(w5anfinwgt0) strata(w#varstrat) || _n, weight(an2015_wt_r) svy: meglm *[outcome predictors]*, family(bernoulli) link(logit) || spid: /* specified as linear model*/ svyset w#varunit, weight(w5anfinwgt0) strata(w#varstrat) || _n, weight(an2015_wt_r) svy: meglm *[outcome predictors]* || spid:
SAS	Not available
R	Not available
Notes	where the data are pooled across rounds an2011_wt_r = (round-specific analytic weight for 2011 cohort: wanfinwgt0 / w1anfinwgt0) an2015_wt_r = (round-specific analytic weight for 2015 cohort: wanfinwgt0 / w5anfinwgt0) multi-level models account for SP-level effect

How do I specify the NHATS design in analyses of time until a non-recurrent event?

Analysts may be interested in time to a specific event (e.g., death or first disability episode) for a specified cohort. In this case, survival analysis can be used.

Analysts should first select a target population; for example they may select the 2011 cohort or the 2015 cohort, depending on the research question. They may then organize their data file with either one row per respondent (wide) or multiple rows per respondent (long). The latter approach allows the inclusion of time-varying covariates. In both examples, the variable *time* indicates time to the event, and the variable *censor* indicates whether this event occurs (a value of "1") or it is censored (a value of "0"). Note that the examples provided use the initial cohort weight (e.g. 2011 for the 2011 cohort or 2015 for the 2015 cohort) and as such do not explicitly take into account loss to follow-up that may occur at subsequent rounds.

Stata	svyset w#varunit [pw=w#anfinwgt0], strata(w#varstrat) stset time, failure(censor) svy: stcox *[predictors]* /*multiple records per respondent with time-varying covariates added*/ svyset w#varunit [pw=w#anfinwgt0], strata(w#varstrat) stset time, failure(censor) id(spid) svy: stcox *[predictors]*
SAS	proc surveyphreg; weight w#anfinwgt0; cluster w#varunit; strata w#varstrat; model time*censor(0) = *[predictors]*; run;
R	library(survey) #need this line only once per session nhats.dsgn <- svydesign(id=~w#varunit, strata=~w#varstrat, weights=~wanfinwgt0, data= *[data frame name]*, nest=TRUE) model.cox <- svycoxph(Surv(time,censor) ~ $[x_1 + x_2 + \ldots + x_{10}]$, design=nhats.dsgn) summary(model.cox)
Notes	where # = 1 for 2011 cohort, and # = 5 for 2015 cohort

How do I specify the NHATS design for analyses focused on outcomes for persons who experience a particular type of event over a specified time period?

Analyses of this type often start by identifying all persons who experienced the event of interest (in any round). Often these analyses use an index event (e.g. first hospitalization, first surgery) so sample persons are not represented more than once. But different individuals may first qualify in different rounds.

We recommend that analyses of this type use either the 2011 cohort or the 2015 cohort to form the sample of interest. While it might seem possible to increase the sample by using the 2011 cohort and

adding persons who were newly included at the 2015 replenishment, the NHATS was not designed to support this approach. Persons newly included in the 2015 replenishment are not representative of persons 65+ by themselves. The weights (and design variables) in 2015 create a nationally representative sample only when the new and continuing sample members are analyzed together.

We further recommend that analysts use the round-specific analytic weight (where the round number is the round they experience the event for the first time) and design variables (psu, strata, which are not round specific). Analysts may also choose to use a subpop command with a variable indicating whether they experienced this event set equal to 1. Such an approach will yield analyses that represent adults ages 65 and older in 2011 (or 2015) who first experience the event of interest over the time period of interest.

How do I specify the NSOC design when using the NSOC III longitudinal file?

To analyze an outcome that takes place in NSOC III among those who participated in NSOC II and III, analysts can do the following. First, create one observation per caregiver by linking the NSOC III longitudinal file to NSOC II. Then specify the cluster variable as c7varunit (if you want to account for geographic clustering of NHATS SPs) or as spid (if you want to account for correlations among caregivers to an SP). Weights from the NSOC III longitudinal sample should be used in order to account for loss to follow up between NSOC II and III.

Stata	svyset c7varunit [pweight=lw7cgfinwgt0], strata(c7varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight lwcgfinwgt0; cluster c7varunit; strata c7varstrat; *[model statements]*; run;
R	library(survey) #need this line only once per session lnsoc.dsgn <- svydesign(id=~c7varunit, strata=~c7varstrat, weights=~lwcgfinwgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	accounts for geographic clustering of observations

Stata	svyset spid [pweight=lw7cgfinwgt0], strata(c7varstrat) svy: *[stata procedures]*
SAS	*[sas procedure]*; weight lwcgfinwgt0; cluster spid; strata c7varstrat; *[model statements]*; run;
R	library(survey) #need this line only once per session lnsoc.dsgn2 <- svydesign(id=~spid, strata=~c7varstrat, weights=~lwcgfinwgt0, data= *[data frame name]*, nest=TRUE) *[model or other statement]*
Notes	accounts for clustering of cg observations within SPs

References

DeMatteis, Jill, Freedman, Vicki A., and Kasper, Judith D. 2016. National Health and Aging Trends Study Round 5 Sample Design and Selection. NHATS Technical Paper #16. Baltimore: Johns Hopkins University School of Public Health.

Freedman, Vicki A., Spillman, Brenda C., and Kasper, Judith D. 2016. Making National Estimates with the National Health and Aging Trends Study. NHATS Technical Paper #17. Johns Hopkins University School of Public Health. Available at www.NHATS.org.

Montaquila, Jill, Freedman, Vicki A., Edwards, Brad, and Kasper, Judith D. 2012. National Health and Aging Trends Study Round 1 Sample Design and Selection. NHATS Technical Paper #1. Baltimore: Johns Hopkins University School of Public Health.