

The Behavioral Risk Factor Surveillance System (BRFSS)

Complex Sampling Weights and
Preparing 2018 BRFSS Module Data for Analysis

July 2019



Overview

The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based telephone survey that collects data on a number of health outcomes, health-related risk behaviors, use of preventive services, and chronic conditions from noninstitutionalized adults who reside in each of the states and participating US territories. In 2011, BRFSS changed its data collection procedures, structure, and weighting methodology so the traditional landline telephone-based data set could include data from participants using cell phones. The BRFSS uses a core set of questions and gives states the option to include multiple modules that focus on specific health issues. Not all optional modules are collected by all of the states, and states may opt to collect module data from just subsets of their survey participants. Core questions are used during all interviews. Researchers using BRFSS data should conduct analyses with complex sampling procedures; they also should appropriately stratify and weight the data in their work. Potential bias resulting from selection probabilities and noncoverage among segments of the population can be reduced through weighting.

Researchers conducting analysis of variables from the core-only section should use the variable _LLCPWT for weighting.

BRFSS has created this document to guide users analyzing variables from 2018 BRFSS optional modules and/or combinations of module and core variables. The 2018 BRFSS data sets include data that respondents provided by landline telephone or cell phone. Data users should note that newer weighting procedures are likely to affect trend lines when comparing BRFSS data collected before and after 2011; because of these changes, users are advised NOT to make direct comparisons with pre-2011 data, and instead, begin new trend lines with that year.

Data users should become familiar with the information presented in this document prior to performing analyses. More information about the changes to the 2018 BRFSS is available in the [2018 Data Comparability Report](#).

Weighting BRFSS Core Data

BRFSS has grouped the 2018 data into four data sets:

- 2018 BRFSS Questionnaire data (Combined Landline Telephone and Cellular Telephone)

And three versions

- 2018 BRFSS Combined Landline Telephone and Cellular Telephone Version 1
- 2018 BRFSS Combined Landline Telephone and Cellular Telephone Version 2
- 2018 BRFSS Combined Landline Telephone and Cellular Telephone Version 3

Researchers should understand that the BRFSS data they want to analyze was collected by asking core or module questions on the annual questionnaire. Users, therefore, may need to use data from up to 4 different data sets, described below.

Use the Combined Landline Telephone and Cell phone data if:

- > The questions corresponding to the data you need are found only in the core section **or**
- > The questions are used in common modules asked on both the landline survey and cellular telephone survey.

(Note: See below, **Using BRFSS Data from Multiple Data Sets** for more about common modules.)

Also:

Use the Version 1, Version 2, and/or Version 3 Combined Landline Telephone and Cell Phone data sets from states collecting multiple version questionnaires.

In all cases, the variable **_STSTR** should be used for stratification, and the variable **_PSU** should be used for clustering, in complex sampling analyses. The following table displays the description of the data, names of the data sets, and the variable names of the final weight.

Data description	Data set name	Final weight variable name
Combined Landline Telephone and Cellular Telephone	LLCP2018	_LLCPWT
Combined Landline Telephone and Cellular Telephone Version 1	LLCP18V1	_LCPWTV1
Combined Landline Telephone and Cellular Telephone Version 2	LLCP18V2	_LCPWTV2
Combined Landline Telephone and Cellular Telephone Version 3	LLCP18V3	_LCPWTV3

Researchers conducting analysis of variables from the *core-only* section should use the variable **_LLCPWT** for weighting. The variable **_STSTR** should be used for stratification, and the variable **_PSU** should be used as a primary sampling unit.

The example below uses the arthritis variable, HAVARTH3, and the physically healthy days variable, PHYSHLTH, in the core to demonstrate using the common weight for Louisiana. Examples are given in SAS and R.

Example SAS Code:

```
libname in '\BRFSS\'; /* Set the location of the BRFSS data set */

proc surveymeans      /* Calculate average number of physical healthy days */
  data = in.llcp2018 (where = (_state = 22));
  weight _llcpwt; /* Use common weight because question is from the core section */
  strata _ststr;
  cluster _psu;
  var physhlth;
run;

proc surveyfreq /* Calculate percentage of population with arthritis */
  data = in.llcp2018 (where = (_state = 22));
  weight _llcpwt; /* Use common weight because question is from the core section */
  strata _ststr;
  cluster _psu;
  table havarth3;
run;
```

Example R Code:

```
# Install the survey package if it is not already installed
install.packages("survey")

# Call the library for the current R session
library(survey)

# Read in BRFSS data
load("\BRFSS\BRFSS.rdata")

# Subset the data for Louisiana
BRFSS <- BRFSS[BRFSS$state == 22,]

# Set options for allowing a single observation per stratum
options(survey.lonely.psu = "adjust")

# Create survey design
brfssdsgn <- svydesign(
  id=~1,
  strata = ~ststr,
  weights = ~llcpwt,
  data = BRFSS)

# calculate average number of physical healthy days
svymean(~physhlth, # Variable to analyze
         brfssdsgn,
         na.rm = TRUE)

# calculate percent in each arthritis category
svymean(~factor(havarth3),
         brfssdsgn, na.rm = TRUE)
```

Using BRFSS Data from Multiple Data Sets

Due to the complex nature of state-based data collection processes, users may have to create a data set that fits their research needs. The following examples illustrate how to prepare module data for analysis when states have collected module data in a variety of ways.

Prior to conducting analyses, users should always review all relevant documents for the given data year (e.g., the [overview](#), [codebook](#), [module list](#), etc.). Users should also double-check FIPS state codes in each data set to avoid duplication.

Individual states may have chosen to use a number of optional modules, depending on each state's needs. Individual states may have chosen to divide their samples and use different modules in the subsamples that were distinguished by the version of the surveys. Modules that appeared in each version of a state's questionnaire are called common modules. Each state that collected common modules did so on all its versions of combined landline telephone and cell phone.

Prior to using optional module data, users must identify the states that collected the data of interest and determine which questionnaire version the state used, if any. Check the 2018 document that lists [modules by state](#).

Example 1: Cognitive Decline Module

The example below uses the Cognitive Decline module to demonstrate how to combine and reweight data from multiple data sets for analysis. This module was selected because states collected the data either as a common module in the combined landline telephone and cell phone or by splitting the sample and offering the module on one of the versions of the data collection. For ease in using the 2018 optional modules, please refer to the two tables published separately with the annual data set: Modules by State and States by Module. In addition to showing which states used which modules, these tables list names and descriptions of the data sets and versions, and weight variables from the corresponding data sets. If users want to analyze variables from certain modules, these two tables are a good place to start. The following information has been copied from the two module documents and show segments of the tables:

Modules by State

State	Description	Dataset	Data weight	Module(s)
Maine	Combined Land Line and Cell Phone data	LLCP2018	_LLCPWT	Childhood Asthma Prevalence, Diabetes, Pre-Diabetes, Random Child Selection
	Combined Land Line and Cell Phone data	LLCP18V1	_LCPWTV1	Cognitive Decline
	Combined Land Line and Cell Phone data	LLCP18V2	_LCPWTV2	E-Cigarettes, Lung Cancer Screening

States by Module

Module	Description	Dataset	Data weight	State(s)
Cognitive Decline	Combined Land Line and Cell Phone data	LLCP2018	_LLCPWT	New Jersey, Oregon, Pennsylvania
	Combined Land Line and Cell Phone data	LLCP18V1	_LCPWTV1	Maine
	Combined Land Line and Cell Phone data	LLCP18V2	_LCPWTV2	Michigan, New York

- 1) Search Cognitive Decline under the [States by Module](#) table. There are 6 states that collected data on this module for 2018.
 - a. Three states collected this data as a common module, New Jersey (FIPS state code=34), Oregon (FIPS state code=41) and Pennsylvania (FIPS state code=42)
 - We need to extract data for New Jersey, Oregon and Pennsylvania from LLCP2018.
 - b. Maine (FIPS state code=23) used version 1 of this module.
 - We therefore need to extract data for Maine from LLCP18V1.
 - c. Michigan (FIPS state code=26) and New York (FIPS state code=36) used version 2 of this module.
 - We therefore need to extract data for Michigan and New York from LLCP18V2.
- 2) Obtain data sets:
 - To download the Combined Landline Telephone and Cellular Telephone data set (LLCP2018):
 - Go to the data [webpage](#), and choose the 2018 data set. From there, read all of the 2018 documents including the [overview](#), [codebook](#), and the [module list](#). The data set is available in ASCII and SAS transport formats. Record layout, formats, and SAS code are included to read in the data.
 - To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 1 data set:
 - Go to the [webpage](#) with Combined Landline and Cellular Telephone Survey Multiple Questionnaire Version Data. There are 3 separate SAS datasets corresponding to Questionnaire Version 1, Questionnaire Version 2, and Questionnaire Version 3. Choose Version 1: LLCP18V1.
 - To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 2 data set:
 - Follow the same step above for Version 1, but instead, choose Version 2: LLCP18V2.
- 3) Generate a new and uniform final weight variable from each of the data sets and combine data into a working data set. This step will require users to:
 - >keep all the states that collect the module data in each of the data sets,
 - >rename the corresponding weight variable to a new and consistent weight variable, and
 - >combine all the data sets into one that contains the renamed weight variable. Details are shown in the SAS syntax below.

Example SAS Code:

libname in '\BRFSS\';

*Extract state data from landline and cell phone combined data;
*Begin with the main dataset, where Cognitive Decline was used as a common module in the landline and cellular telephone samples;
*Select states where the Cognitive Decline module was a common module;

data llcp;
set in.llcp2018 (**where**=(_state in (34, 41, 42))); *New Jersey, Oregon & Pennsylvania;

*Rename weight variable _LLCPWT to _FINALWT to be consistent across new data sets;
rename _llcpwt=_finalwt;
run;

*Extract data from Questionnaire VI;
data llcpv1;

*Select states where the Cognitive Decline module was collected on questionnaire version 1;
set in.llcp18v1 (**where**=(_state in (23))); *Only Maine used Version 1;

*Rename weight variable _LCPWTV1 to _FINALWT to be consistent across new data sets;
rename _lcpwtv1=_finalwt;
run;

*Questionnaire Version 2: Extract states from Questionnaire V2;
data llcpv2;

*Select states where the Cognitive Decline module was collected on questionnaire version 2;
set in.llcp18v2 (**where**=(_state in (26, 36))); *Michigan & New York;

*Rename final weight variable _LCPWTV2 to _FINALWT to be consistent across new data sets;
rename _lcpwtv2=_finalwt;
run;

*Combine all 3 data sets together;
data Cognitv;
set llcp llcpv1 llcpv2;
run;

Example R Code:

```
# Begin with the main data set for states that used the Cognitive Decline module  
# as a common module.
```

```
# FIPS state codes: 34 (New Jersey), 41 (Oregon), 42 (Pennsylvania)  
llcp <- llcp2018[llcp2018$state %in% c(34, 41, 42), ]
```

```
# Rename the final weight variable for the main data set  
llcp$finalwt <- llcp$llcpwt
```

```
# Choose the states that used the Cognitive Decline module in version 1  
# FIPS state codes: 23 (Maine)
```

```
llcpv1 <- llcp18v1[llcp18v1$state %in% c(23), ]
```

```
# Rename the final weight variable for the version 1 data set  
llcpv1$finalwt <- llcpv1$lcpwtv1
```

```
# Choose the states that used the Cognitive Decline module in version 2
```

```
# FIPS state codes: 26 (Michigan) 36 (New York)
llcpv2 <- llcp18v2[llcp18v2$state %in% c(26,36), ]
```

```
# Rename the final weight variable for the version 2 data set
llcpv2$finalwt <- llcpv2$lcpwtv2
```

```
# Combine all data sets
```

```
cognitv <- rbind(llcp, llcpv1, llcpv2)
```

Example 2: E-Cigarettes Module

The example below uses the E-Cigarettes Module to demonstrate how to combine and reweight data from multiple data sets. This module was selected for illustration because it covers multiple questionnaire versions.

- 1) Search E-Cigarette module under the [States by Module](#) table. There are a total 38 states that used the E-Cigarettes optional module in 2018.
 - 33 states including Guam used this module as a common version: Alaska, Arkansas, Connecticut, Delaware, Florida, Georgia, Guam, Hawaii, Idaho, Indiana, Iowa, Louisiana, Maryland, Massachusetts, Minnesota, Mississippi, Missouri, Montana, Nebraska, New Hampshire, New York, North Carolina, North Dakota, Ohio, Oregon, Rhode Island, South Dakota, Tennessee, Texas, Utah, Virginia, Wisconsin, Wyoming
 - We need to extract data for these 33 states from the data set LLCP2018.
 - One state, Colorado (FIPS state code=8), used version 1 of the module
 - We need to extract data for Colorado from LLCP18V1.
 - Three states, Kansas (FIPS state code=20), Maine (FIPS state code=23) & Michigan (FIPS state code=26), used version 2 of the module.
 - We need to extract data for Kansas, Maine & Michigan from LLCP18V2
 - Michigan (FIPS state code=26), also used version 3 of this module
 - We need to extract data for Michigan from LLCP18V3.
- 2) If a state collects data on a module on **more** than one questionnaire versions, there is an extra step to creating the appropriate analysis weight. For instance, Michigan collected the E-Cigarette module on both Questionnaire Versions 2 and 3. In this case, we should check and compare the sample size in each of the versions of the data set and decide how to calculate the final weight by multiplying by a proportion of the whole.

Here is an example of SAS code that shows Michigan's use of the E-Cigarette module: For those states that used more than one version of the module, the sample size for each version should be checked so that the weights can be adjusted proportionally. The SAS code below subsets data for version 2 and version 3 of the module for Michigan. The result shows that it has 3359 observations in version 2 and 3394 observations in version 3 of the combined landline telephone and cell-phone data for 2018. Since the number of observations in both versions are very similar, the analysis weight (_LCPWTW2) for Version 2 and the analysis weight (_LCPWTW3) for Version 3 can simply be divided by 2.

If somehow the number of observations are quite different, for instance, version 2 has about 1500, and version 3 has 3000, then we will calculate the proportion for each version and use the calculated proportion times _LCPWT2 and _LCPWT3 respectively to get the adjusted weight for each version. The calculation process will be:

- 1) Get the sum of total observations for both versions of the module $1500+3000=4500$
- 2) The proportion for version 2: $1500/4500 = 0.33$
- 3) The proportion for version 3: $3000/4500 = 0.67$
- 4) The adjusted weight for version 2 is $_LCPWT2 * 0.33$, and the adjusted weight for version 3 is $_LCPWT3 * 0.67$

Example SAS code:

```
Data MIv2;  
Set brfss.LLCP18V2(keep=_state where=(_state=26));  
Title "Michigan Sample Size, Questionnaire Version 2, Combined Landline and Cell Phone";  
run;
```

SAS log shows: There were 3359 observations read from the data set BRFSS.LLCP18V2 WHERE _state=26;

```
Data MIv3;  
Set brfss.LLCP18V3(keep=_state where=(_state=26));  
Title "Michigan Sample Size, Questionnaire Version 3, Combined Landline and Cell Phone";  
run;
```

SAS log shows: There were 3394 observations read from the data set BRFSS.LLCP18V3, WHERE _state=26;

3) Obtain data sets:

- To download the Combined Landline Telephone and Cellular Telephone data set (LLCP2018):
 - Go to the data [webpage](#), and choose the 2018 data set. From there, read all of the 2018 documents including the [overview](#), [codebook](#), and the [module list](#). The data set is available in ASCII and SAS transport formats. A record layout, format syntax, and SAS code are included to read in the raw data.
- To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 1 dataset:
 - Go to the webpage with Combined Landline and Cellular Telephone Survey Multiple Questionnaire Version Data. There are 3 separate SAS data sets corresponding to Questionnaire Version 1, Questionnaire Version 2, and Questionnaire Version 3. Choose Version 1: LLCP18V1.
- To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 2 data set:
 - Follow the same step above for Version 1, but instead, choose Version 2: LLCP18V2.
- To download the Combined Landline Telephone and Cellular Telephone Questionnaire Version 3 data set:

- Follow the same step above for Version 1, but instead choose Version 3: LLCP18V3.

Generate a new and uniform final weight variable from each of the data sets, and combine data into a working data set. This step will require users to:

- a. Keep all of the states that collect the module data in each of the data sets;
- b. Rename the corresponding weight variable to a consistent weight variable; and
- c. Combine all the data sets into one that contains the consistent weight variable.

Details are shown in the SAS code below.

Example SAS code:

libname in '\BRFSS\';

*Begin with the main dataset, where E-Cigarette was used as a common module in the combined landline and cellular telephone samples;

*for these 33 states: Alaska, Arkansas, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Indiana, Iowa, Louisiana, Maryland, Massachusetts, Minnesota, Mississippi, Missouri, Montana, Nebraska, New Hampshire, New York, North Carolina, North Dakota, Ohio, Oregon, Rhode Island, South Dakota, Tennessee, Texas, Utah, Virginia, Wisconsin, Wyoming, Guam;

```

data llcp;
set in.llcp2018 (where=(_state in (2,5,9,10,12,13,15,16,18,19,22,24,25,27,28,29,
30,31,33,36,37,38,39,41,44,46,47,48,49,51,55,56,66)));
rename _llcpwt=_finalwt; *Rename weight variable to be consistent across new data sets;
run;

*Extract data for states that used Questionnaire V1;
data llcpv1;
set in.llcp18v1 (where=(_state = 8); *For Colorado only (2018);
Renam_lcpwtv1=_finalwt; *Rename weight variable to be consistent across new data sets;
run;

*Extract data for Kansas(22), Maine(23) & Michigan(26) that used Questionnaire V2;
data llcpv2;
set in.llcp18v2 (where=(_state in(20,23,26)));
Renam_lcpwtv2=_finalwt; *Rename weight variable to be consistent across new data sets.

*Combine all datasets together;
data ECig;
set llcp llcpv1 llcpv2 llcpv3;
run;

```

Example R code:

```
# Determine number of observations in the version 2 data set for Michigan
nrow(llcp18v1[llcp18v1$state == 26, ])

# Determine number of observations in the version 2 data set for Michigan
nrow(llcp18v2[llcp18v2$state == 26, ])

# Begin with the main data set for states that used the E-Cigarette module
# as a common module, 33 states including Guam.
llcp <- llcp2018[llcp2018$state %in% c(2,5,9,10,12,13,15,16,18,19,22,24,25,27,28,29,
30,31,33,36,37,38,39,41,44,46,47,48,49,51,55,56,66), ]

# Rename the final weight variable for the main data set
llcp$finalwt <- llcp$llcpwt

# Extract data for Colorado (FIPS code=8) state that used version 1 of the E-Cigarette module
llcpv1 <- llcp18v1[llcp18v1$state %in% c(8), ]

# Rename the final weight variable for the version 1 data set
llcp$finalwt <- llcp$lcpwtv1

# Extract data for states (Kansas, Maine & Michigan) that used the Version 2 of the E-Cigarette module
llcpv2 <- llcp18v2[llcp18v2$state %in% c(20, 23,26), ]

# Adjust weight for Michigan by dividing version 2 weight by 2 because Michigan collected
# version2 and version 3 of the E-Cigarette module
# and the sample size in the two versions is roughly equal
llcpv2$finalwt <- ifelse(llcpv2$state == 26, lcpwtv2/2, lcpwtv2)

# Extract data for Michigan state that used the Version 3 of the E-Cigarette module
llcpv3 <- llcp18v3[llcp18v3$state == 26, ]

# Rename the final weight variable for the version 3 data set
llcpv3$finalwt <- llcpv3$lcpwtv3/2

# Combine the data sets
ecig <- rbind(llcp, llcpv1, llcpv2, llcpv3)
```

Combining Multiple Years of BRFSS Data

There are times we need to analyze data from multiple years, such as calculating the average prevalence of a variable in core modules for 2017 and 2018. In this case we need to combine data from 2017 and 2018. The weight variables for 2017 and 2018 need be adjusted proportionally for the combined data set. There are two ways to adjust the weight variable. The sample size for each year helps to determine how we should adjust them. If the sample sizes are very similar, we can simply divide the originally weight variable by 2. If they are different, we need to adjust the weight variable proportionally by following these steps:

- 1) Find out the sample size for 2017 and 2018
- 2) Sum up the total sample size for 2017 and 2018
- 3) Find the proportion for 2017 (use the sample size for 2017 divided by the combined sample size for 2017 & 2018)
- 4) Find the proportion for 2018 (use the sample size for 2018 divided by the combined sample size for 2017 & 2018)
- 5) The adjusted weight for 2017 is ($_LLCPWT(2017)$ * proportion for 2017)
- 6) The adjusted weight for 2018 is ($_LLCPWT(2018)$ * proportion for 2018)

The example below uses the 2017 and 2018 data from New York to demonstrate how to combine and reweight data on the core section from multiple years.

We have found that in 2017 the sample size for New York is 12,249, whereas in 2018 the sample size is 35,767. The sample size for 2018 is much larger than that in 2017. We determined that their weight needs to be adjusted proportionally. Following the steps listed above, we find:

- 1) 2017 sample size is 12,249, 2018 is 35,767
- 2) The combined sample size is 48,016
- 3) The proportion for 2017 is $12,249/48,016 = .26$;
- 4) The proportion for 2018 is $35,767/48,016 = .74$
- 5) The adjusted weight for 2017 is ($_LLCPWT(2017)$ * .26)
- 6) The adjusted weight for 2018 is ($_LLCPWT(2018)$ * .74)

Obtain data sets:

- To download the 2017 Combined Landline Telephone and Cellular Telephone Questionnaire data set (LLCP2017):
 - Go to the data [webpage](#), and choose the 2017 data set. From there, read all of the 2017 documents including the [overview](#), [codebook](#), and the [module list](#). The data set is available in ASCII and SAS transport formats. A record layout, format syntax, and SAS code are included to read in the raw data.
- To download the 2018 Combined Landline Telephone and Cellular Telephone Questionnaire data set (LLCP2018):
 - Go to the data [webpage](#), and choose the 2018 dataset. From there, read all of the 2018 documents including the [overview](#), [codebook](#), and the [module list](#). The data set is available in ASCII and SAS transport formats. A record layout, format syntax, and SAS code are included to read in the raw data.

Generate a new and uniform final weight variable (the adjusted weight variables) from each of the data sets, and combine data into one working data set. This step will require users to:

- a. Rename the corresponding weight variable to a consistent weight variable; and
- b. Combine all the data sets into one data set that contains the consistent weight variable.

Details are shown in the SAS syntax below.

Data users should note that new weighting procedures are likely to affect trend lines when comparing BRFSS data collected before and after 2011; because of these changes, users are advised NOT to make direct comparisons with pre-2011 data, and instead, to begin new trend lines with that year.

Example SAS Code:

```

/* Check totals for each year */
proc sql;
title "Check Totals in 2017";
  select count(*)
  from in16.llcp2017
  where _state = 36;

title "Check Totals in 2018";
  select count(*)
  from in18.llcp2018
  where _state = 36;
quit;

/* Create weighted dataset with the adjusted weight variable (finalwt) for each year */

data combineYears;
  set in17.llcp2017 (in = a)
    in18.llcp2018 (in = b);
  if a then finalwt = _llcpwt * (.26);
  if b then finalwt = _llcpwt * (.74);
  where _state = 36;
run;

```

Example R Code:

```

# Subset 2017 and 2018 data for New York
NY17 <- BRFSS17[BRFSS17$state == 36, ]
NY18 <- BRFSS18[BRFSS18$state == 36, ]

# Count number of observations in each year
nrow(NY17)
nrow(NY18)

# Create a final weight based on proportion of observations in each year
NY17$finalwt <- NY17$llcpwt * (.26)
NY18$finalwt <- NY18$llcpwt * (.74)

# Combine the data sets and reweight
NY1718 <- rbind(NY17, NY18)

```