

Behavioral Risk Factor Surveillance System (BRFSS)

Complex Sampling Weights and Preparing 2019 BRFSS Module Data for Analysis

July 2020



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 2019 BRFSS optional modules and/or combinations of module and core variables. The 2019 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 2019 BRFSS is available in the [2019 Data Comparability Report](#).

Weighting BRFSS Core Data

BRFSS has grouped the 2019 data into four data sets:

- 2019 BRFSS Questionnaire data (Combined Landline Telephone and Cellular Telephone)

And three versions

- 2019 BRFSS Combined Landline Telephone and Cellular Telephone Version 1
- 2019 BRFSS Combined Landline Telephone and Cellular Telephone Version 2
- 2019 BRFSS Combined Landline Telephone and Cellular Telephone Version 3

Researchers should understand that 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 cellphone survey (*Note: See below, Using BRFSS Data from Multiple Data Sets for more about common modules.*)
- Use Version 1, Version 2, and/or Version 3 Combined Landline Telephone and Cell Phone data sets if states collected multiple version questionnaires.

In all cases, researchers conducting complex sampling analyses from the *core-only* section should use the variable **_LLCPWT** for weighting, **_STSTR** for stratification, and the variable **_PSU** for primary sampling unit, or clustering for short. 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	LLCP2019	_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

The examples below demonstrate how to use **_LLCPWT**, **_STSTR** & **_PSU** when analyzing variable **HAVARTH4** (*Ever told you had some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia*) and **PHYSHLTH** (*how many days during the past 30 days was your physical health not good?*) in core section using 2019 Louisiana dataset. Examples are given in both SAS and R.

Example in SAS:

```
LIBNAME IN '\BRFSS\';      /* Use LIBNAME to set the location of the BRFSS data set */

PROC SURVEYMEANS          /* Calculate average number of physical healthy days */
  DATA = IN.LLCP2019 (WHERE=( _STATE= 22)); /* Fips code for Louisiana is 22 */
  WEIGHT _LLCPWT;        /* Use _LLCPWT because question is from the core section */
  STRATA _STSTR;
  CLUSTER _PSU;
  VAR PHYSHLTH;
RUN;

PROC SURVEYFREQ;         /* Calculate percentage of population with arthritis */
  DATA = IN.LLCP2019(WHERE= ( _STATE = 22));
  WEIGHT _LLCPWT;       /* Use common weight because question is from the core section */
  STRATA _STSTR;
```

```
CLUSTER _PSU;  
TABLE HAVARTH4;  
RUN;
```

Example in R:

```
# 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  
# Variable to analyze  
svymean(~physhlth,  
  brfssdsgn,  
  na.rm = TRUE)  
  
# calculate percent in each arthritis category  
svymean(~factor(havarth4),  
  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 using optional module data, users should always review all relevant documents for the given data year, such as [Overview](#), [Codebook](#), [Comparability of Data](#), and [Modules by State](#) in particular to identify states that collected the data of interest and determine which questionnaire version states used. Remember to 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 also 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.

Example 1: Adverse Childhood Experience

The example below uses the Adverse Childhood Experience 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 2019 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, 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 to show segments of the tables:

Modules by State

State	Description	Dataset	Data weight	Module(s)
Ohio	Combined Land Line and Cell Phone data	LLCP2019	_LLCPWT	Childhood Asthma Prevalence, Health Care Access, Random Child Selection, Sexual Orientation and Gender Identity (SOGI)
	Combined Land Line and Cell Phone data	LLCP19V1	_LCPWTV1	Adverse Childhood Experience, Breast and Cervical Cancer Screening, Cognitive Decline, Colorectal Cancer Screening, Diabetes, Family Planning, Home/ Self-measured Blood Pressure, Pre-Diabetes
	Combined Land Line and Cell Phone data	LLCP19V2	_LCPWTV2	Caregiver

States by Module

Module	Description	Dataset	Data weight	State(s)
Adverse Childhood Experience	Combined Land Line and Cell Phone data	LLCP2019	_LLCPWT	Alabama, Delaware, Florida, Indiana, Iowa, Michigan, Mississippi, Missouri, New Mexico, North Dakota, Pennsylvania, Rhode Island, South Carolina, Tennessee, Virginia, West Virginia, Wisconsin
	Combined Land Line and Cell Phone data	LLCP19V1	_LCPWTV1	Ohio
	Combined Land Line and Cell Phone data	LLCP19V2	_LCPWTV2	Kansas, Oklahoma
	Combined Land Line and Cell Phone data	LLCP19V3	_LCPWTV3	New York

1) **Search Adverse Childhood Experience** under the [States by Module](#) table. There are 21 states collected data on this module for 2019. Seventeen states used the common version of the module, 1 state used version 1, 2 states used version 2 and 1 state used version 3 of the module. See detailed information below:

- a. Seventeen states collected this data as a common module therefore we need to extract data for these states from LLCP2019. The seventeen states are: Alabama, Delaware, Florida, Indiana, Iowa, Michigan, Mississippi, Missouri, New Mexico, North Dakota, Pennsylvania, Rhode Island, South Carolina, Tennessee, Virginia, West Virginia, Wisconsin
- b. Ohio used version 1 of this module, we therefore need to extract data for Ohio from LLCP19V1.
- c. Kansas, Oklahoma used version 2 of this module, we need to extract data for these 2 states from LLCP19V2.
- d. New York used version 3 of this module, therefore we need to extract data from LLCP19V3 for New York.

2) **Obtain data sets:**

- To download the Combined Landline Telephone and Cellular Telephone data set (LLCP2019):
 - Go to the data [webpage](#), and choose the 2019 data set. From there, read all of the 2019 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, 2, and 3. Choose Version 1: [LLCP19V1](#).
- 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: [LLCP19V2](#).
- 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: [LLCP19V3](#)

3) **Generate a working dataset:**

Create a new and uniform final weight variable from each version of the data sets downloaded in the step above, combine data into one 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 which contains the renamed weight variable. Details are shown in the SAS syntax below.

Example in SAS:

```
LIBNAME IN '\BRFSS\';
```

```
*Extract state data from landline and cell phone combined data;
```

```
*Begin with the main dataset, where Adverse Childhood Experience was used as a common module in the landline and cellular telephone samples;
```

```
DATA LLCP; /*Select states where the Adverse Childhood Experience module was a common module; */
SET IN.LLCP2019 (WHERE=(STATE IN (1,10,12,18,19,26,28,29,35,38,42,44,45,47,51,54,55)));
RENAME _LLCPWT=_FINALWT;
RUN;
```

```
DATA LLCPV1;
SET IN.LLCP19V1 (WHERE=(STATE IN (39))); *state of OH(Fips=39) used version 1;
RENAME _LCPWTV1=_FINALWT;
RUN;
```

```
DATA LLCPV2;
SET IN.LLCP19V2 (WHERE=(STATE IN (20,40))); *Kansas, Oklahoma used version 2;
RENAME _LCPWTV2=_FINALWT;
RUN;
```

```
DATA LLCPV3;
SET IN.LLCP19V3 (WHERE=(STATE IN (36))); *New York used version 3;
RENAME _LCPWTV3=_FINALWT;
RUN;
```

```
DATA ADVERSE; *Combine all 3 datasets which can be used to analyze the data;
SET LLCP LLCPV1 LLCPV2 LLCPV3;
RUN;
```

Example in R:

```
# Begin with the main data set for states that used the Adverse Childhood Experience as a common module.
```

```
# FIPS state codes for the 17 states that used common version of the module
```

```
llcp <- llcp2019[llcp2019$state %in% c(1,10,12,18,19,26,28,29,35,38,42,44,45,47,51,54,55), ]
```

```
# Rename the final weight variable for the main data set
```

```
llcp$finalwt <- llcp$llcpwt
```

```
# Choose the states that used version 1 of the module – OH(fips=39)
```

```
llcpv1 <- llcp19v1[llcp19v1$state %in% c(39), ]
```

```

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

# Choose the states that used version 2 of the module – Kansas, Oklahoma
# FIPS state codes: 20 (Kansas), 40(Oklahoma)
llcpv2 <- llcp19v2[llcp18v2$state %in% c(20,40), ]

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

# Choose the states that used version 3 of the module – New York
# FIPS state codes: 36(New York)
llcpv3 <- llcp19v3[llcp19v3$state %in% c(36) ]

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

# Combine all data sets
adverse <- rbind(llcp, llcpv1, llcpv2,llcpv3)

```

Example 2: Family Planning

The example below uses the Family Planning Module to demonstrate how to combine and reweight data from multiple data sets.

- 1) **Search Family Planning module** under the [States by Module](#) table. There are total 38 states used the Family Planning optional module in 2019.
 - 34 states used the common version of this module, therefore we need to extract data for these 34 states from the dataset LLC2019. These states are: Alabama, Arizona, Arkansas, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas, Louisiana, Maryland, Massachusetts, Minnesota, Mississippi, Missouri, Montana, New Mexico, North Carolina, Oregon, Pennsylvania, Puerto Rico, Rhode Island, South Carolina, South Dakota, Tennessee, Utah, Virginia, West Virginia, Wisconsin, Wyoming
 - 2 states, Ohio and Oklahoma, used version 1 of the module, therefore we need to extract data from LLC19V1 for these 2 states
 - 2 states, Nebraska and New York used version 2 of the Family Planning module and therefore we need to extract data for these 2 states from LLC19V2.
- 2) **Obtain data sets:**
 - To download the Combined Landline Telephone and Cellular Telephone data set (LLC2019):
 - Go to the data [webpage](#), and choose the 2019 data set. From there, read all of the 2019 documents including the [Overview](#), [Codebook](#), and the [States by Module](#). 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, 2, and 3. Choose Version 1: [LLCP19V1](#).
 - 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: [LLCP19V2](#).
- 3) **Generate a working dataset:** Create 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 renamed weight variable. Details are shown in the code below

Example code in SAS:

```
LIBNAME IN '\BRFSS\';
```

**Begin with the main dataset, where Family Planning was used as a common module in the combined landline and cellular telephone samples for 34 states: Alabama, Arizona, Arkansas, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas, Louisiana, Maryland, Massachusetts, Minnesota, Mississippi, Missouri, Montana, New Mexico, North Carolina, Oregon, Pennsylvania, Puerto Rico, Rhode Island, South Carolina, South Dakota, Tennessee, Utah, Virginia, West Virginia, Wisconsin, Wyoming;*

```
DATA LLCP;
```

```
SET IN.LLCP2019 (WHERE=( _STATE in  
(1,4,5,9,10,12,13,15,16,17,18,19,20,22,24,25,27,28,29,30,35,37,41,42,44,45,46,47,49,51,54,55,56,72)));
```

```
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.LLCP19V1 (WHERE=( _state IN (39,40)); *Ohio & Oklahoma;
```

```
RENAME _LCPWTV1 = _FINALWT;      *Rename weight variable to be consistent across new data sets;  
RUN;
```

**Extract data for states that used Questionnaire V2;*

```
DATA LLCPV2;
```

```
SET IN.LLCP19V2 (WHERE=( _STATE IN (31,36)); *Nebraska & New York;
```

**Rename weight variable to be consistent across new data sets.*

```
RENAME _LCPWTV2 = _FINALWT;  
RUN;
```

**Combine all datasets together;*

```
data FAMPLN;  
SET LLCP LLCPV1 LLCPV2;  
RUN;
```

Dataset FAMPLN is ready to be analyzed

Example in R:

```
# Begin with the main data set for 34 states that used the Family Planning module as a common module  
llcp <- llcp2019[llcp2019$state %in%  
c(1,4,5,9,10,12,13,15,16,17,18,19,20,22,24,25,27,28,29,30,35,37,41,42,44,45,46,47,49,51,54,55,56,72), ]
```

```
# Extract Questionnaire V1 data for Ohio & Oklahoma  
nrow(llcp19v1[llcp19v1$state %in% c(39,40), ])
```

```
# Extract Questionnaire V2 data for Nebraska & New York  
nrow(llcp19v2[llcp19v2$state %in% c(31,36), ])
```

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

```
# Rename the final weight variable for the version 1 data set  
llcpv1$finalwt <- llcpv1$llcpwtv1
```

```
# Rename the final weight variable for the version 2 data set  
llcpv2$finalwt <- llcpv2$llcpwtv2
```

```
# Combine the data sets and it's ready for analysis  
fampln <- rbind(llcp, llcpv1, llcpv2)
```

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 sections for 2018 and 2019. In this case we need to combine data from 2018 and 2019. The weight variable for 2018 and 2019 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 original weight variable by 2. If they are different, we need to adjust the weight variable proportionally by follow these steps:

- 1) Find out the sample size for 2018 and 2019
- 2) Sum up the total sample size for 2018 and 2019
- 3) Find the proportion for 2018 (use the sample size for 2018 divided by the combined sample size for 2018 & 2019)
- 4) Find the proportion for 2019 (use the sample size for 2019 divided by the combined sample size for 2018 & 2019)
- 5) The adjusted weight for 2018 is ($_LLCPWT(2018) * \text{proportion for 2018}$)

6) The adjusted weight for 2019 is ($_LLCPWT(2019) * \text{proportion for 2019}$)

The example below uses the 2018 and 2019 data from New York to demonstrate how to combine and reweight data on the core section from multiple years.

After checking the sample size for New York for 2018 and 2019, we found that it is 35,767 for 2018 and 14,232 for 2019. The sample size for 2018 is much larger than that for 2019. We determined that their weight needs to be adjusted proportionally. Following the steps listed above to find the adjusted weight for New York for 2018 & 2019.

- 1) The sample size is 35,767 for 2018, and 14,232 for 2019
- 2) The combined sample size is 49,999
- 3) The proportion for 2018 is $35,767/49,999 = .72$
- 4) The proportion for 2019 is $14,232/49,999 = .28$;
- 5) The adjusted weight for 2018 is ($_LLCPWT(2018) * .72$)
- 6) The adjusted weight for 2019 is ($_LLCPWT(2019) * .28$)

Obtain data sets:

- To download the 2018 Combined Landline Telephone and Cellular Telephone Questionnaire 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 [Modules By State](#). 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 2019 Combined Landline Telephone and Cellular Telephone Questionnaire data set (LLCP2019):
 - Go to the data [Webpage](#), and choose the 2019 dataset. From there, read all of the 2019 documents including the [Overview](#), [Codebook](#), and the [Modules By State](#). 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:
 - Rename the corresponding weight variable to a consistent weight variable; and
 - Combine all the data sets into one data set that contains the consistent weight variable.

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

Reweighting Example in SAS:

```
Libname D18 "\BRFSS2018";  
Libname D19 "\BRFSS2019";
```

```
/* Check sample size for each 2018 & 2019 */
```

```
DATA NY18;
```

```
SET D18.LLCP2018(WHERE=(STATE=36));  
RUN;  
/* NOTE: There were 35767 observations read from the data set D18.LLCP2018 WHERE STATE=36*/
```

```
DATA NY19;  
SET D19.LLCP2019(WHERE=(STATE=36));  
RUN;  
/* NOTE: There were 14232 observations read from the data set D19.LLCP2019 WHERE STATE=36*/  
  
/* Create weighted dataset with the adjusted weight variable (FINALWT) for each year */
```

```
DATA COMBINEYRS;  
SET D18.LLCP2018 (IN=A)  
D19.LLCP2019 (IN=B);  
WHERE STATE=36;  
IF A THEN FINALWT = LLCPTWT * .72);  
IF B THEN FINALWT = LLCPTWT * .28);  
RUN;
```

Reweighting Example in R:

```
# Subset 2018 and 2019 data for New York  
NY18 <- llcp2018[llcp2018$state == 36, ]  
NY19 <- llcp2019[llcp2019$state == 36, ]  
  
# Count number of observations in each year  
nrow(NY18)  
nrow(NY19)  
  
# Create a final weight based on proportion of observations in each year  
NY18$finalwt <- NY18$llcpwt * (.72)  
NY19$finalwt <- NY19$llcpwt * (.28)  
  
# Combine the data sets and reweight  
NY1819 <- rbind(NY18, NY19)
```