# 1999 to 2015NATIONAL SURVEYs ON DRUG USE AND HEALTH

Small Area Estimation Dataset: State Small Area Estimates, by survey year, outcome, state, and age group

Substance Abuse and Mental Health Services Administration  
Center for Behavioral Health Statistics and Quality

Rockville, Maryland 20857

October 2017

1999 to 2015 NATIONAL SURVEYS ON DRUG USE AND HEALTH

SMALL AREA ESTIMATION DATASET: STATE SMALL AREA ESTIMATES, BY SURVEY YEAR, OUTCOME, STATE, AND AGE GROUP

Contract Nos. HHSS283201000003C and HHSS283201300001C

Project Nos. 0212800.001.110.002.003, 0213984.004.109.002, and 0214839.001.008

For questions about this codebook and data file, please e-mail [Peter.Tice@samhsa.hhs.gov](mailto:Peter.Tice@samhsa.hhs.gov).

Prepared for Substance Abuse and Mental Health Services Administration, Rockville, Maryland

Prepared by RTI International, Research Triangle Park, North Carolina

Data File Name: States\_saes\_final.sas7bdat

October 2017

Recommended Citation: Center for Behavioral Health Statistics and Quality. (2017). *1999 to 2015 National Surveys on Drug Use and Health: Small Area Estimation Dataset: State Small Area Estimates, by Survey Year, Outcome, State, and Age Group*. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Table of Contents

Chapter Page

[1. Introduction 1](#_Toc462649026)

[2. Variables on the Dataset 3](#_Toc462649027)

[2.1 OUTCOME 3](#_Toc462649028)

[2.2 OUTNAME 6](#_Toc462649029)

[2.3 STNAME 6](#_Toc462649030)

[2.4 STATE 6](#_Toc462649031)

[2.5 AREA 7](#_Toc462649032)

[2.6 AGEGRP 7](#_Toc462649033)

[2.7 PYEAR 8](#_Toc462649034)

[2.8 PYEARNM 8](#_Toc462649035)

[2.9 POP 9](#_Toc462649036)

[2.10 NSEL 9](#_Toc462649037)

[2.11 NCOMP 9](#_Toc462649038)

[2.12 WTINTRR 9](#_Toc462649039)

[2.13 BSAE 9](#_Toc462649040)

[2.14 LOW\_SAE 10](#_Toc462649041)

[2.15 UP\_SAE 10](#_Toc462649042)

[2.16 STE\_SAE 10](#_Toc462649043)

[2.17 GROUP 10](#_Toc462649044)

[2.18 EST\_TOTAL 11](#_Toc462649045)

[2.19 LOW\_TOTAL 11](#_Toc462649046)

[2.20 UP\_TOTAL 11](#_Toc462649047)

[2.21 STE\_TOTAL 11](#_Toc462649048)

[2.22 GEN\_CORR 12](#_Toc462649049)

Appendix

[A Comparison of Small Area Estimates A-1](#_Toc462649179)

## 1. Introduction

This dataset contains state-level small area estimates, associated confidence intervals, and other key statistics related to state-level, model-based estimates of certain key substance use and mental health outcomes from the 1999 to 2015 National Surveys on Drug Use and Health (NSDUHs).[[1]](#footnote-1) State-level NSDUH small area estimates have been published annually by pooling 2 years of NSDUH data since 1999. Hence, this dataset contains small area estimates from the following pooled years of NSDUHs:[[2]](#footnote-2)

* 1999-2000,
* 2000-2001,
* 2002-2003,
* 2003-2004,
* 2004-2005,
* 2005-2006,
* 2006-2007,
* 2007-2008,
* 2008-2009,
* 2009-2010,
* 2010-2011,
* 2011-2012,
* 2012-2013,
* 2013-2014, and

2014-2015.

Note that pooled NSDUH small area estimates were not produced using 2001 and 2002 combined data because the 2002 data differed from the data from the 1999 through 2001 surveys. In 2002, several changes were introduced to the survey. Incentives of $30 were given to respondents for the first time in order to address concerns about the response rates. Other changes included a change in the survey name (i.e., from NHSDA to NSDUH), new data collection quality control procedures, and a shift from the 1990 decennial census to the 2000 census as a basis for population count totals and to calculate any census-related predictor variables that are used in small area estimation. An unanticipated result of these changes was that the prevalence rates for 2002 were in general substantially higher than those for 2001—higher than could be attributable to the usual year-to-year trends—and thus are not comparable with estimates for 2001 and prior years. Therefore, the 2002 NSDUH was established as a new baseline for both the national and the state estimates. For more details, refer to Section A.2 of the "2011-2012 NSDUH: Guide to State Tables and Summary of Small Area Estimation Methodology" at <https://www.samhsa.gov/data/>.

The purpose of producing this dataset is to have available, in one location, all of the state-level small area estimates, from the earliest (1999-2000) to the latest (2014-2015). This dataset does not provide any new information, but it does offer all of the available information in a more user-friendly format (i.e., as a SAS dataset instead of as HTML and PDF-Web tables). This will allow users to analyze the data for a specific state, year, or outcome by subsetting the file. Also, if a user is interested in testing differences between two state estimates or testing differences between a state's estimates across time, all of the information needed is on this dataset. Appendix A provides details on how to use the data to conduct these tests.

This dataset contains 98,880 observations. The records are included at the year × outcome × age group × state level. There are 22 variables on this dataset, and the variable descriptions for each of these 22 variables are provided in Chapter 2.

## 2. Variables on the Dataset

This chapter describes the 22 variables included on this 1999 to 2015 National Survey on Drug Use and Health (NSDUH) state small area estimation (SAE) dataset, which is sorted by PYEAR, STATE, OUTCOME, and AGEGRP. Note that when an estimate is not available, it will either show up as "." in the SAS dataset or that observation will not be included. For example, GROUP is not defined for national and regional estimates (i.e., when STNAME = National, Northeast, Midwest, South, and West). Thus, for those observations, the value for GROUP is SAS missing (i.e., GROUP = .). Alternatively, for 1999-2000, regional estimates are not defined and hence are not included in this dataset (i.e., there is no observation for PYEARNM = 1999-2000 and STNAME = Northeast or Midwest or South or West).

### 2.1 OUTCOME

"OUTCOME" is the name for the key substance use or mental health measures. Unless otherwise noted, the outcome name matches the variable name on the NSDUH analytic file. Note that not all outcomes are available for all years. For more information about which outcomes are available in which years, see Table C.15 of the "2014-2015 NSDUH: Guide to State Tables and Summary of Small Area Estimation Methodology" at <https://www.samhsa.gov/data/>. The values of "OUTCOME" are as follows:

* ABODALC: alcohol dependence or abuse in the past year (available for 2000-2001 and beyond, but small area estimates[[3]](#footnote-3) were not produced for this outcome in 1999-2000),
* ABODILAL: dependence or abuse of illicit drugs or alcohol in the past year (available for 2000-2001 through 2013-2014, but small area estimates were not produced for this outcome in 1999-2000 or in 2014-2015),
* ABODILL: illicit drug dependence or abuse in the past year (available for 2000-2001 through 2013-2014, but small area estimates were not produced for this outcome in 1999-2000 or in 2014-2015),
* ALCMON: alcohol use in the past month (available for all years),
* AMIYR: any mental illness in the past year (available for 2008-2009 and beyond, but data are not available in previous years),[[4]](#footnote-4)
* ANLYR: nonmedical use of pain relievers in the past year (available for 2002-2003 through 2013-2014, but small area estimates were not produced for this outcome in prior years or in 2014-2015),
* BNGALC: binge alcohol use in the past month (available for all years except 2014-2015),[[5]](#footnote-5)
* CIGMON: cigarette use in the past month (available for all years),
* COCYR: cocaine use in the past year (available for all years),
* DEPNDALC: alcohol dependence in the past year (available for 2000-2001 and beyond, but small area estimates were not produced for this outcome in 1999-2000),
* DEPNDILL: illicit drug dependence in the past year (available for 2000-2001 through 2013-2014, but small area estimates were not produced for this outcome in 1999-2000 or in 2014-2015),
* GRSKHTRY: perceptions of great risk from trying heroin once or twice (available for 2013-2014 only),
* HERYR: heroin use in the past year (available for 2013-2014 and beyond),
* IEMMON: illicit drug use other than marijuana in the past month (available for all years except 2014-2015),
* INCIDENCE: average annual rate of first use of marijuana (available for all years),[[6]](#footnote-6)
* MDE: had at least one major depressive episode in the past year (i.e., depression) (available for 2005-2006 and beyond),[[7]](#footnote-7)
* MRJMON: marijuana use in the past month (available for all years),
* MRJYR: marijuana use in the past year (available for 2002-2003 and beyond, but small area estimates were not produced for this outcome in prior years),
* RISKALC: perceptions of great risk from having five or more drinks of an alcoholic beverage once or twice a week (available for all years except 2014-2015),[[8]](#footnote-8)
* RISKCIG: perceptions of great risk from smoking one or more packs of cigarettes per day (available for all years except 2014-2015),[[9]](#footnote-9)
* RISKMJ: perceptions of great risk from smoking marijuana once a month (available for all years except 2014-2015),[[10]](#footnote-10)
* SMIYR: serious mental illness in the past year (available for 2008-2009 and beyond, with the question first added to NSDUH in 2008),[[11]](#footnote-11)
* SPD\_L: serious psychological distress in the past year based on the long-form questionnaire (available for 2002-2003 and 2003-2004),[[12]](#footnote-12)
* SPD\_S: serious psychological distress in the past year based on the short-form questionnaire (available for 2004-2005),[[13]](#footnote-13)
* SUITHKYR: had serious thoughts of suicide in the past year (available for 2008-2009 and beyond, with the question first added to NSDUH in 2008),[[14]](#footnote-14)
* SUMMON: illicit drug use in the past month (available for all years except 2014-2015),
* TOBMON: tobacco product use in the past month (available for all years),
* TXGAPALC: needing but not receiving treatment at a specialty facility for alcohol use in the past year (available for 2002-2003 through 2013-2014, but small area estimates were not produced for this outcome in prior years or in 2014-2015),
* TXGPILAL: needing but not receiving treatment at a specialty facility for substance use in the past year (available for 2010-2011 through 2013-2014),
* TXNOSPEC: needing but not receiving treatment at a specialty facility for illicit drug use in the past year (available for 2000-2001 through 2013-2014, but small area estimates were not produced for this outcome in 1999-2000 or in 2014-2015),
* TXREC3: received mental health services in the past year (available for 2010-2011 and beyond),
* U\_ALCMON: underage (among persons aged 12 to 20) alcohol use in the past month (available for 2002-2003 and beyond, but small area estimates were not produced for this outcome in prior years),[[15]](#footnote-15) and

U\_BNGALC: underage (among persons aged 12 to 20) binge alcohol use in the past month (available for 2002-2003 through 2013-2014, but small area estimates were not produced for this outcome in prior years or in 2014-2015).[[16]](#footnote-16)

### 2.2 OUTNAME

"OUTNAME" is the label for each OUTCOME variable. These labels are included in the OUTCOME descriptions in Section 2.1.

### 2.3 STNAME

"STNAME" is the name of the census region or state. Its values are as follows:

* *National*: United States;
* *Northeast*: Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont;
* *Midwest*: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin;
* *South*: Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia;
* *West*: Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming; and

the 51 names of the states (Alabama, Alaska, …Wisconsin, and Wyoming) and the District of Columbia.

### 2.4 STATE

A federal information processing standards (FIPS) code has been assigned for each "STATE," including the District of Columbia.[[17]](#footnote-17) The FIPS code takes on values from -5 to -1 to denote the national estimates and estimates for the four census regions:

* -5: National,
* -4: Northeast,
* -3: Midwest,
* -2: South,
* -1: West,
* 1: Alabama,
* 2: Alaska,
* 4: Arizona,
* 5: Arkansas,
* 6: California,
* 8: Colorado,
* 9: Connecticut,
* 10: Delaware,
* 11: District of Columbia,
* 12: Florida,
* 13: Georgia,
* 15: Hawaii,
* 16: Idaho,
* 17: Illinois,
* 18: Indiana,
* 19: Iowa,
* 20: Kansas,
* 21: Kentucky,
* 22: Louisiana,
* 23: Maine,
* 24: Maryland,
* 25: Massachusetts,
* 26: Michigan,
* 27: Minnesota,
* 28: Mississippi,
* 29: Missouri,
* 30: Montana,
* 31: Nebraska,
* 32: Nevada,
* 33: New Hampshire,
* 34: New Jersey,
* 35: New Mexico,
* 36: New York,
* 37: North Carolina,
* 38: North Dakota,
* 39: Ohio,
* 40: Oklahoma,
* 41: Oregon,
* 42: Pennsylvania,
* 44: Rhode Island,
* 45: South Carolina,
* 46: South Dakota,
* 47: Tennessee,
* 48: Texas,
* 49: Utah,
* 50: Vermont,
* 51: Virginia,
* 53: Washington,
* 54: West Virginia,
* 55: Wisconsin, and

56: Wyoming.

Sorting (in increasing order) this 1999 to 2015 dataset by state will ensure that the national estimates come first, then the census region estimates, then the estimates for all of the states. Note that regional estimates (i.e., census regions) are not available for 1999-2000.

### 2.5 AREA

"AREA" is a numeric variable for a recode of STATE to differentiate between the national, regional, and state estimates. Its values are as follows:

* 0: national (STATE = -5),
* 1: the four census regions (STATE = ‑4, ‑3, ‑2, or -1), and

2: the 50 states and the District of Columbia (STATE ≥ 1).

### 2.6 AGEGRP

"AGEGRP" is a numeric age group variable. Its values are as follows:

* 0: persons aged 12 or older,
* 1: persons aged 12 to 17,
* 2: persons aged 18 to 25,
* 3: persons aged 26 or older,
* 4: persons aged 18 or older, and

5: persons aged 12 to 20 (available only for U\_ALCMON and U\_BNGALC).

Note that estimates for persons aged 12 to 20 (for alcohol use and binge alcohol use) are not available for 1999-2000 and 2000-2001.

### 2.7 PYEAR

"PYEAR" is a numeric variable for a pooled pair of survey years. Its values are as follows:[[18]](#footnote-18)

* 1: 1999-2000,
* 2: 2000-2001,
* 3: 2002-2003,
* 4: 2003-2004,
* 5: 2004-2005,
* 6: 2005-2006,
* 7: 2006-2007,
* 8: 2007-2008,
* 9: 2008-2009,
* 10: 2009-2010,
* 11: 2010-2011 published estimates,
* 12: 2010-2011 updated estimates,
* 13: 2011-2012,
* 14: 2012-2013,
* 15: 2013-2014, and

16: 2014-2015.

### 2.8 PYEARNM

"PYEARNM" is a variable for a pooled pair of survey years. Its values are as follows:

* 1999-2000,
* 2000-2001,
* 2002-2003,
* 2003-2004,
* 2004-2005,
* 2005-2006,
* 2006-2007,
* 2007-2008,
* 2008-2009,
* 2009-2010,
* 2010-2011 published estimates,
* 2010-2011 updated estimates,
* 2011-2012,
* 2012-2013,
* 2013-2014, and

2014-2015.

Note that for 2010-2011, two sets of state-level small area estimates were produced. The 2010-2011 published estimates are available at <https://www.samhsa.gov/data/> and were based on predictors and weights that used the 2000 census as its base. The 2010-2011 updated estimates were developed using new predictors and new weights obtained from the 2010 census. Note that the updated 2010-2011 estimates were only produced to be compared with the published estimates and not to replace them.

The 1999-2000 and 2000-2001 small area estimates used the 1990 census as a base source for obtaining population controls for survey weights and also used the 1990 census as the source of predictors. The 2002-2003 to 2010-2011 published small area estimates used the 2000 census as their source, and the 2010-2011 updated estimates through the 2013-2014 small area estimates were based on the 2010 census.

### 2.9 POP

The next four variables (i.e., POP, NSEL, NCOMP, and WTINTRR) represent population and sample characteristics and are not dependent on the outcome measure. They are provided for each PYEAR by AGEGRP by STATE combination and are the same for each outcome. The NSEL, NCOMP, and WTINTRR are based on the respondent's age at screening. Note that this age can differ from the respondent's age at the time of the interview. Thus, the values here for the individual age groups may differ slightly from the numbers reported elsewhere based on the respondent's age at the time of the interview (note that the numbers for those aged 12 or older will be same).

"POP" is a variable for the estimated number of persons in the population averaged across the 2 survey years.

### 2.10 NSEL

"NSEL" is a variable for the number of persons selected for the survey summed over the 2 survey years. Additional information is provided in the POP variable description.

### 2.11 NCOMP

"NCOMP" is a variable for the number of respondents who completed the survey summed over the 2 survey years (i.e., the pooled sample size over 2 survey years). Additional information is provided in the POP variable description.

### 2.12 WTINTRR

"WTINTRR" is a variable for the weighted interview response rate averaged across the 2 survey years (expressed as a proportion). Additional information is provided in the POP variable description.

### 2.13 BSAE

The next three variables (i.e., BSAE, LOW\_SAE, and UP\_SAE) represent the small area estimates (i.e., prevalence rates of substance use and mental health outcomes and corresponding confidence intervals [CIs]) that are published each year. These estimates are expressed as proportions and can be multiplied by 100 to be expressed as percentages. For some years, the national CIs were not published, but they have been included in this dataset (any exceptions are noted below). For more information about point estimates, SAE methodology, exact benchmarking, and CIs, see the "2010-2011 National Survey on Drug Use and Health: Guide to State Tables and Summary of Small Area Estimation Methodology" at <https://www.samhsa.gov/data/>.

"BSAE" is a variable for the benchmarked (i.e., adjusted) small area estimate where the national small area estimate is benchmarked to be equal to the national design-based estimate. For more details on exact benchmarking, refer to Section B.6 of the "2011-2012 National Surveys on Drug Use and Health: Guide to State Tables and Summary of Small Area Estimation Methodology" at <https://www.samhsa.gov/data/>.

The state and census region benchmarked small area estimates are based on the hierarchical Bayes estimation approach. The national estimates (prevalence rates and corresponding CIs), however, are design-based estimates. Note that exact benchmarking was introduced in 2002; hence, for the estimates prior to that (i.e., 1999-2000 and 2000-2001), the national design-based estimate and the national small area estimate will not be the same. The estimate is still called BSAE, but for 1999-2000 and 2000-2001, the national estimate included in this dataset is a Bayes model-based estimate, not the national design-based estimate. For all other years, the national design-based estimate is provided for STNAME = "National."

### 2.14 LOW\_SAE

"LOW\_SAE" is a variable for the 95 percent lower Bayesian CI associated with BSAE. For the national estimates, design-based CIs are provided. Note that the national CIs provided for 1999-2000 and 2000-2001 are the model-based Bayesian CIs (because these estimates are not benchmarked, the national estimate is a Bayes model-based estimate).

### 2.15 UP\_SAE

"UP\_SAE" is a variable for the 95 percent upper Bayesian CI associated with BSAE. For the national estimates, design-based CIs are provided. Note that the national CIs provided for 1999-2000 and 2000-2001 are the model-based Bayesian CIs (because these estimates are not benchmarked, the national estimate is a Bayes model-based estimate).

### 2.16 STE\_SAE

"STE\_SAE" is a variable for the standard error (SE) associated with BSAE. For the national estimates, design-based SEs are included. Note that STE\_SAE values are not provided for 1999-2000 and 2000-2001.

Note also that the CIs mentioned above (LOW\_SAE and UP\_SAE) are not calculated as a symmetric interval directly from this STE\_SAE. For details on this, see Section A.4 of the "2011-2012 National Surveys on Drug Use and Health: Guide to State Tables and Summary of Small Area Estimation Methodology" at <https://www.samhsa.gov/data/>.

### 2.17 GROUP

"GROUP" is a variable for a map group with a range from 1 to 5. The BSAEs are sorted and grouped into five quintiles to produce maps. Group 1 represents estimates that are the lowest estimates, and group 5 represents the largest estimates. Because national and regional estimates are not included in these groupings, the variable GROUP has missing values for these records. Only the 50 states and the District of Columbia are grouped, and maps are produced based on these five groupings.

State estimates (which are based on a survey-weighted hierarchical Bayes [SWHB] estimation approach) fall into one of five quintiles according to their ranking. Because 51 states were ranked for each measure, the middle quintile was assigned to 11 states, and the remaining quintiles were assigned 10 states each. In some cases, a "quintile" could have more or fewer states than desired because two (or more) states had the same estimate (to two decimal places). When such ties occurred at the "boundary" between two quintiles, all of the states with the same estimate were conservatively assigned to the lower quintile. The map groups were created separately for each pair year (PYEAR), measure (OUTCOME), and age group (AGEGRP). For more details, refer to the "2011-2012 National Survey on Drug Use and Health National Maps of Prevalence Estimates, by State" at <https://www.samhsa.gov/data/>. Note that for most outcomes, maps were not published for persons aged 18 or older, but map groups have been included in this file.

### 2.18 EST\_TOTAL

The next three variables (i.e., EST\_TOTAL, LOW\_TOTAL, and UP\_TOTAL) represent the small area estimate counts rounded to the nearest thousand and corresponding CIs that are published each year. For some years, the national CIs were not published, but they have been included in this dataset unless otherwise noted.

"EST\_TOTAL" is a variable for the average number (in thousands) of persons having the outcome of interest across the 2 survey years, as follows:

EST\_TOTAL = (BSAE × POP / 1,000), rounded.

The totals for persons aged 18 or older are calculated as sums of 18 to 25 totals and the 26 or older totals. Similarly, the 12 or older totals are produced as the sum of the 12 to 17, 18 to 25, and the 26 or older totals. The totals for the national and census regions are the sum of the totals for all states that make up the given area.

### 2.19 LOW\_TOTAL

"LOW\_TOTAL" is a variable for the 95 percent lower CI associated with EST\_TOTAL, as follows:

LOW\_TOTAL = (LOW\_SAE × POP / 1,000), rounded.

### 2.20 UP\_TOTAL

"UP\_TOTAL" is a variable for the 95 percent upper CI associated with EST\_TOTAL, as follows:

UP\_TOTAL = (UP\_SAE × POP / 1,000), rounded.

### 2.21 STE\_TOTAL

"STE\_TOTAL" is a variable for the SE associated with EST\_TOTAL, as follows:

STE\_TOTAL = (STE\_SAE × POP / 1,000), rounded.

Note that STE\_TOTAL values are not provided for 1999-2000 and 2000-2001; however, they can be calculated using the above formula.

### 2.22 GEN\_CORR

"GEN\_CORR" is a generalized correlation that can be used for statistical testing of percentages between nonoverlapping years. No generalized correlations are available for OUTCOME = HERYR, GRSKHTRY, TXGPILAL, or TXREC3. Specifically, for a given state, outcome, and age group, it is the correlation between the log odds of the estimate at time period 1 and the log odds of the estimate at time period 2 where the two time periods do not overlap. See below for additional information.

For each nonmental health-related outcome measure (OUTCOME) by state (STNAME) by age group (AGEGRP) combination, the generalized correlation is an average of seven correlations:

* 2002-2003 versus 2007-2008,
* 2002-2003 versus 2008-2009,
* two sets of 2002-2003 versus 2009-2010,[[19]](#footnote-19)
* 2002-2003 versus 2010-2011,
* 2002-2003 versus 2012-2013, and

2002-2003 versus 2013-2014.

For the mental health outcome measures, the correlations used to calculate the generalized correlations were different. For the OUTCOME = MDE by state (STNAME) by age group (AGEGRP) combination, the generalized correlation is an average of eight correlations:

* 2005-2006 versus 2007-2008,
* 2005-2006 versus 2008-2009,
* 2005-2006 versus 2009-2010,
* 2005-2006 versus 2010-2011,
* 2005-2006 versus 2011-2012,
* 2005-2006 versus 2012-2013,
* 2006-2007 versus 2009-2010, and

2008-2009 versus 2010-2011.

For the other mental health outcomes (OUTCOME = AMIYR, SMIYR, SUITHKYR) by state (STNAME) by age group (AGEGRP) combinations, the generalized correlation is an average of six correlations:

* 2008-2009 versus 2010-2011,
* 2008-2009 versus 2011-2012,
* 2008-2009 versus 2012-2013,
* 2009-2010 versus 2011-2012,
* 2009-2010 versus 2012-2013, and

2010-2011 versus 2012-2013

Each of these sets of correlations was produced by simultaneously fitting 4 years of NSDUH data separately for each outcome measure. For example, to produce correlations between the 2002-2003 and 2007-2008 state estimates for past month marijuana use, four age groups (12 to 17, 18 to 25, 26 to 34, and 35 or older) by two time periods (2002-2003 and 2007-2008), or eight subpopulation-specific models, were fitted, each with its own set of fixed and random effects. In this case, the general covariance matrices for the state and within-state random effects were 8 × 8 matrices corresponding to the eight element (age group × time period) vectors of random effects. Note that the survey-weighted, Bernoulli-type log likelihood employed in the SWHB methodology was appropriate for this simultaneous model because the eight age group × year subpopulations were nonoverlapping. The correlation between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat was approximated by the correlation calculated using the posterior distributions of the natural logarithm of pi 1 sub s and a divided by 1 minus pi 1 sub s and a and the natural logarithm of pi 2 sub s and a divided by 1 minus pi 2 sub s and a from the simultaneous model.

Note that these generalized correlations are same for each year (2002-2003 through 2013-2014) and are not defined for 1999-2000 and 2000-2001. The mental health generalized correlations are defined for all years where the mental health outcome is defined. These generalized correlations are meant to be used to calculate between-year differences for a given state and only for *nonoverlapping* years, such as 2004-2005 versus 2008-2009. The correlations between overlapping years are in general higher than these nonoverlapping generalized correlations. However, if an analyst wants to test any differences in state estimates between any 2 nonoverlapping years, such as 2011-2012 versus 2009-2010, or 2008-2009 versus 2004-2005, or any earlier years going back to 2002-2003, these correlations can be used. The national estimates are direct estimates, so the correlations for these are zero.

Section A.2 of Appendix A describes the methodology for conducting these tests. Tests of differences in state estimates for overlapping years (e.g., 2010-2011 vs. 2011-2012, or 2008-2009 vs. 2009-2010) can be found on the SAMHSA website. Note that generalized correlations are not available for OUTCOME = SPD. Because SPD is defined for only 3 years, generalized correlations were not produced.

Appendix A: Comparison of Small Area Estimates

A.1 Comparison of Two Small Area Estimates (within a Given Year)

This section describes a method for determining whether differences between two state estimates within a given time period are statistically significant. This procedure can be used for any two state estimates representing the same age group (e.g., young adults aged 18 to 25, AGEGRP = 2) and time period (e.g., 2009-2010, PYEAR = 10).

Let pi 1 sub a and pi 2 sub a denote the 2009-2010 age group-*a* specific prevalence rates for two different states, state 1 and state 2, respectively. The difference between pi 1 sub a and pi 2 sub a is defined in terms of the log-odds ratio (lor sub a) as opposed to the simple difference because the posterior distribution of lor sub a is closer to Gaussian than the posterior distribution of the simple difference pi 1 sub a minus pi 2 sub a. The lor sub a is defined as The log-odds ratio, lor sub a, is defined as the natural logarithm of the ratio of two quantities. The numerator of the ratio is pi 2 sub a divided by 1 minus pi 2 sub a. The denominator of the ratio is pi 1 sub a divided by 1 minus pi 1 sub a., where ln denotes the natural logarithm. The *p* value is computed to the test the null hypothesis of no difference (i.e., Pi 1 sub a is equal to pi 2 sub a. or equivalently, lor sub a equals zero). An estimate of lor sub a is given by The estimate of the log-odds ratio, lor hat sub a, is defined as the natural logarithm of the ratio of two quantities. The numerator of the ratio is p 2 sub a divided by 1 minus p 2 sub a. The denominator of the ratio is p 1 sub a divided by 1 minus p 1 sub a, where p 1 sub a is the 2009-2010 state estimate for state s1 and age group-a, and p 2 sub a is the 2009-2010 state estimate for state s2 and age group-a for a particular outcome of interest. where p 1 sub a and p 2 sub a are the 2009-2010 state estimates (BSAE).[[20]](#footnote-20) To compute the variance of the estimate of the log-odds ratio, lor hat sub a that is, the variance v of the estimate of the log-odds ratio, lor hat sub a let Theta 1 hat be defined as the ratio of p 1 sub a and 1 minus p 1 sub a and Theta 2 hat be defined as the ratio of p 2 sub a and 1 minus p 2 sub a then The variance v of the estimate of the log-odds ratio, lor hat sub a, is a function of three quantities: q1, q2, and q3. It is expressed as the sum of q1 and q2 minus q3. Quantity q1 is the variance v of the natural logarithm of Theta 1 hat, quantity q2 is the variance v of the natural logarithm of Theta 2 hat, and quantity q3 is 2 times the covariance between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat. where the covariance between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat denotes the covariance between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat This covariance is defined in terms of the associated correlation as follows:

The covariance between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat is equal to the correlation between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat multiplied by the square root of the product of the variance v of the natural logarithm of Theta 1 hat and the variance v of the natural logarithm of Theta 2 hat..

The quantities variance v of the natural logarithm of Theta 1 hat and variance v of the natural logarithm of Theta 2 hat can be obtained by using the 95 percent Bayesian confidence intervals (CIs), namely (UP\_SAE, LOW\_SAE). For this purpose, let lower sub 1 and upper sub 1 and lower sub 2 and upper sub 2 denote the 95 percent Bayesian CIs for the two states, s sub 1 representing state 1 and s sub 2 representing state 2, respectively. Then

Variance v of the natural logarithm of Theta sub i is equal to the square of quantity q. Quantity q is calculated as the difference between capital U sub i and capital L sub i divided by 2 times 1.96, where i takes values 1 and 2.

where Capital U sub i is the natural logarithm of upper sub i divided by 1 minus upper sub i, and capital L sub i is the natural logarithm of lower sub i divided by 1 minus lower sub i.

For all practical purposes, the correlation between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat is assumed to be negligible; hence, the variance v of the estimate of the log-odds ratio, lor hat sub a can be approximated by the sum of the variance v of the natural logarithm of Theta 1 hat and the variance v of the natural logarithm of Theta 2 hat. The correlation is assumed to be negligible because each state was a stratum in the first level of stratification; therefore, each state sample is selected independently. However, the correlation between the two state estimates is theoretically nonzero because state estimates share common fixed-effect parameters in the small area estimation (SAE) models. Hence, the test statistic quantity z (defined below) might result in a different conclusion in a few cases when the correlation between the state estimates is incorporated in calculating the variance v of the estimate of the log-odds ratio, lor sub a To calculate the p value for testing the null hypothesis of no difference (Log-odds ratio, lor sub a, is equal to zero.), it is assumed that the posterior distribution of log-odds ratio, lor sub a is normal with Mean is equal to the estimate of the log-odds ratio, lor hat sub a. and Variance is equal to the variance v of the estimate of the log-odds ratio, lor hat sub a. With the null value of Log-odds ratio, lor sub a, is equal to zero., the Bayes *p*value or significance levels for the null hypothesis of no difference is The p value is equal to 2 times the probability of realizing a standard normal variate greater than or equal to the absolute value of a quantity z., where capital Z is a standard normal random variate, Quantity z is the estimate of the log-odds ratio, lor hat sub a, divided by the square root of the sum of the variance v of the natural logarithm of Theta 1 hat and the variance v of the natural logarithm of Theta 2 hat., and absolute value of quantity z denotes the absolute value of quantity z This Bayesian significance level (or *p* value) for the null value of log-odds ratio lor
, say log-odds ratio lor sub zero, is defined following Rubin[[21]](#footnote-21) as the posterior probability for the collection of the log-odds ratio lor
values that are less likely or have smaller posterior density d of the log-odds ratio lor than the null (no change) value log-odds ratio lor sub zero. That is, The p value of log-odds ratio lor sub zero is equal to the probability of d of the log-odds ratio lor when it is less than or equal to d of the log-odds ratio lor sub zero.. With the posterior distribution of log-odds ratio lor
 approximately normal, the p value of log-odds ratio lor sub zero is given by the above expression.

**Example.** The 2009-2010 prevalence rates for past month alcohol use among 12 to 17 year olds in Minnesota and North Dakota are shown in the following exhibit. Looking at the two 95 percent Bayesian CIs, it would appear that the Minnesota and North Dakota prevalence rates for past month alcohol use are not statistically different at the 5 percent level of significance because the two Bayesian CIs overlap.

| **STNAME** | **Point Estimate (%) = BSAE** | **95% Bayesian Confidence Interval (%) ( LOW\_SAE, UP\_SAE)** |
| --- | --- | --- |
| Minnesota | 0.1316 | (0.1110, 0.1555) |
| North Dakota | 0.1658 | (0.1425, 0.1920) |

However, in the following discussion, the test based on the quantity z statistic described earlier concludes that they are significantly different at the 5 percent level of significance.

Let p 1 sub a equal 0.1316, lower sub 1 equal 0.1110, upper sub 1 equal 0.1555, p 2 sub a equal 0.1658, lower sub 2 equal 0.1425 upper sub 2 equal 0.1920. Then,

Capital U sub 1 is defined as the natural logarithm of the ratio of 0.1555 and 1 minus 0.1555, which is -1.6921. 

Capital L sub 1 is defined as the natural logarithm of the ratio of 0.1110 and 1 minus 0.1110, which is -2.0806. 

Capital U sub 2 is defined as the natural logarithm of the ratio of 0.1920 and 1 minus 0.1920, which is -1.4371. 

Capital L sub 2 is defined as the natural logarithm of the ratio of 0.1425 and 1 minus 0.1425, which is -1.7947.

The estimate of the log-odds ratio, lor hat sub a, is defined as the natural logarithm of the ratio of two quantities. The numerator of the ratio is p 2 sub a divided by 1 minus p 2 sub a. The denominator of the ratio is p 1 sub a divided by 1 minus p 1 sub a, where p1 sub a is 0.1316 and p 2 sub a is 0.1658. The estimate lor hat sub a is calculated to be 0.2712.

The variance v of the natural logarithm of Theta 1 hat is equal to the square of quantity q. Quantity q is calculated as the difference between capital U sub 1 and capital L sub 1 divided by the product of 2 and 1.96. Here, capital U sub 1 is -1.6921, and capital L sub 1 is -2.0806. Hence, the variance v of the natural logarithm of Theta 1 hat is calculated to be 0.00982.
 
The variance v of the natural logarithm of Theta 2 hat is equal to the square of quantity q. Quantity q is calculated as the difference between capital U sub 2 and capital L sub 2 divided by the product of 2 and 1.96. Here, capital U sub 2 is -1.4371, and capital L sub 2 is -1.7947. Hence, the variance v of the natural logarithm of Theta 2 hat is calculated to be 0.00832.

Quantity z is the estimate of the log-odds ratio, lor hat sub a, divided by the square root of the sum of the variance v of the natural logarithm of Theta 1 hat and the variance v of the natural logarithm of Theta 2 hat, where lor hat sub a is 0.2712, the variance v of the natural logarithm of Theta 1 hat is 0.00982, and the variance v of the natural logarithm of Theta 2 hat is 0.00832. The statistic z is calculated to be 2.0134.

Because the computed absolute value of quantity z is greater than or equal to 1.96 (the critical value of the quantity z statistic), then at the 5 percent level of significance, the hypothesis of no difference (Minnesota prevalence rate = North Dakota prevalence rate) is rejected. Thus, the two state prevalence rates are statistically different. The Bayes *p* value for the null hypothesis of no difference is The Bayes p value or posterior probability of no difference is calculated as 2 times the probability that capital Z is greater than or equal to 2.0134. The p value is equal to 0.044. Hence, this difference would be considered significant at the 5 percent level of significance.

A.2 Comparison of Nonoverlapping Year Small Area Estimates

This section describes a method for determining whether differences between two nonoverlapping time periods (i.e., 2002-2003 and 2011-2012) for a given state are statistically significant. To determine whether the differences between two nonoverlapping state prevalence rates at time period 1 and time period 2 are statistically significant, let pi 1 sub s and a and pi 2 sub s and a denote the prevalence rates at time period 1 and time period 2, respectively, for state-*s* and age group-*a.* The difference between pi 1 sub s and a and pi 2 sub s and a is defined in terms of the log‑odds ratio lor sub s and a as opposed to the simple difference because the posterior distribution of log-odds ratio, lor sub s and a is closer to Gaussian than the posterior distribution of the simple difference Pi 2 sub s and a minus pi 1 sub s and a represents a simple difference between two prevalence rates.  The log-odds ratio, lor sub s and a is defined as

The log-odds ratio, lor sub s and a, is defined as the natural logarithm of the ratio of two quantities. The numerator of the ratio is pi 2 sub s and a divided by 1 minus pi 2 sub s and a. The denominator of the ratio is pi 1 sub s and a divided by 1 minus pi 1 sub s and a.,

where ln denotes the natural logarithm. The *p* value is computed to test the null hypothesis of no change (i.e., Pi 2 sub s and a is equal to pi 1 sub s and a. or equivalently, Log-odds ratio, lor sub s and a, is equal to zero.). An estimate of log-odds ratio, lor sub s and a is given by

The estimate of the log-odds ratio, lor hat sub s and a, is defined as the natural logarithm of the ratio of two quantities. The numerator of the ratio is p 2 sub s and a divided by 1 minus p 2 sub s and a. The denominator of the ratio is p 1 sub s and a divided by 1 minus p 1 sub s and a, where p 1 sub s and a are the state estimates for time period 1 and p2 sub s and are the state estimates for time period 2.

where p 1 sub s and a and p 2 sub s and a are the state estimates (BSAEs) for the 2 years being compared. To compute the variance of the estimate of the log-odds ratio, lor hat sub s and a that is, variance v of the estimate of the log-odds ratio, lor hat sub s and a let Theta 1 hat equal the ratio of p 1 sub s and a and 1 minus p 1 sub s and a and Theta 2 hat equal the ratio of p 2 sub s and a and 1 minus p 2 sub s and a then

Variance v of the estimate of the log-odds ratio, lor hat sub s and a, is a function of three quantities: q1, q2, and q3. It is expressed as the sum of q1 and q2 minus q3. Quantity q1 is the variance v of the natural logarithm of Theta 1 hat, quantity q2 is the variance v of the natural logarithm of Theta 2 hat, and quantity q3 is 2 times the covariance between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat.

where the covariance between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat denotes the covariance between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat This covariance is defined in terms of the associated correlation as follows:

The covariance between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat is equal to the correlation between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat multiplied by the square root of the product of the variance v of the natural logarithm of Theta 1 hat and the variance v of the natural logarithm of Theta 2 hat.,

where Variance v of the natural logarithm of Theta sub i is equal to the square of quantity q. Quantity q is calculated as the difference between capital U sub i and capital L sub i divided by 2 times 1.96, where i takes values 1 and 2. Capital U sub i is the natural logarithm of upper sub i divided by 1 minus upper sub i. Capital L sub i is the natural logarithm of lower sub i divided by 1 minus lower sub i. and the *lower* and *upper* are the 95 percent Bayesian CIs, LOW\_SAE and UP\_SAE.

For the correlation between the natural logarithm of Theta 1 hat and the natural logarithm of Theta 2 hat for an outcome measure by state by age group, the generalized correlation (GEN\_CORR) will be used.

To calculate the *p* value for testing the null hypothesis of no difference (Log-odds ratio lor is equal to zero.), it is assumed that the posterior distribution of log-odds ratio lor is normal with Mean is equal to the estimate of the log-odds ratio, lor hat sub s and a. and Variance is equal to the variance v of the estimate of the log-odds ratio, lor hat sub s and a. With the null value of (Log-odds ratio lor is equal to zero.), the Bayes *p* value or significance levels for the null hypothesis of no difference is The p value is equal to 2 times the probability of realizing a standard normal variate greater than or equal to the absolute value of a quantity z., where capital Z is a standard normal random variate, Quantity z is the estimate of the log-odds ratio, lor hat sub s and a, divided by the square root of the variance v of the estimate of the log-odds ratio, lor hat sub s and a., and absolute value of quantity z denotes the absolute value of quantity z This Bayesian significance level (or *p* value) for the null value of log-odds ratio lor
, say log-odds ratio lor sub zero, is defined following Rubin[[22]](#footnote-22) as the posterior probability for the collection of the log-odds ratio lor
 values that are less likely or have smaller posterior density d of the log-odds ratio lor than the null (no change) value log-odds ratio lor sub zero. That is, The p value of log-odds ratio lor sub zero is equal to the probability of d of the log-odds ratio lor when it is less than or equal to d of the log-odds ratio lor sub zero.. With the posterior distribution of log-odds ratio lor
 approximately normal, the p value of log-odds ratio lor sub zero is given by the above expression.

For overlapping time periods,[[23]](#footnote-23) *p* values are given in published state reports and web documents, and the method described here should *not* be used. Also, because of changes to the survey in 2002, these generalized correlations should not be used to test differences between 1999-2000 small area estimates or 2000-2001 small area estimates and the other small area estimates beyond 2002.

1. Prior to 2002, this survey was referred to as the National Household Survey on Drug Abuse (NHSDA). [↑](#footnote-ref-1)
2. Substate small area estimates are not contained in this file. [↑](#footnote-ref-2)
3. Small area estimates are model-based estimates. For more information, see the "2011-2012 NSDUH: Guide to State Tables and Summary of Small Area Estimation Methodology" at <https://www.samhsa.gov/data/>. [↑](#footnote-ref-3)
4. The variable name for AMIYR on the analytic file is AMIYR\_U. [↑](#footnote-ref-4)
5. The variable name for BNGALC on the analytic file is BINGEDRK. [↑](#footnote-ref-5)
6. INCIDENCE was created using the following variables from the analytic file: MRJFLAG (lifetime marijuana use); INTDATE (interview date); IRMJYFU, IRMJMFU, and IRMJDFU (year, month, and date of first marijuana use, respectively); and MJUSENV (never used marijuana). [↑](#footnote-ref-6)
7. The variable name for MDE on the analytic file is YMDEYR2 for youths aged 12 to 17 and AMDEYR2 for adults aged 18 or older. MDE estimates for youths 12 to 17 are available for 2004-2005 on the Substance Abuse and Mental Health Services Administration (SAMHSA) website. However, those MDE estimates are not included on this file because other age group estimates are not available for 2004-2005. Note that MDE questions were added in 2004 and administered to half of the NSDUH sample. Beginning in 2005, the full sample received the questions. [↑](#footnote-ref-7)
8. RISKALC was created using the following variable from the analytic file: RK5ALWK. [↑](#footnote-ref-8)
9. RISKCIG was created using the following variable from the analytic file: RSKPKCIG. [↑](#footnote-ref-9)
10. RISKMJ was created using the following variable from the analytic file: RSKMJOCC. [↑](#footnote-ref-10)
11. The variable name for SMIYR on the analytic file is SMIYR\_U. [↑](#footnote-ref-11)
12. In the 2002-2003 NSDUH state SAE report, available on the SAMHSA website, this outcome was referred to as "serious mental illness," but it is actually an estimate of SPD. To produce the 2002-2003 small area estimates, the SMI variable from the analytic file was used; for the 2003-2004 small area estimates, the SMI variable was used for 2003, and the SPD\_ADJ variable from the analytic file was used for 2004. Basically, the variables from the analytic file were renamed to SPD\_L, then used in SAE models to produce SPD\_L estimates. In 2004, half of the respondents received the long-form SPD questions, and half of the respondents received the short-form SPD questions. The SPD\_ADJ variable was created by adjusting the short-form responses to match the long-form responses, thus creating a variable that could be combined with the long-form only variable from 2003 (i.e., SMI). [↑](#footnote-ref-12)
13. To produce the 2004-2005 small area estimates, the SPD\_RADJ variable from the analytic file was used for 2004, and the SPD\_UADJ variable from the analytic file was used for 2005. Basically, the variables from the analytic file were renamed to SPD\_S, then used in SAE models to produce SPD\_S estimates. As noted earlier, in 2004, half of the respondents received the long-form SPD questions, and half of the respondents received the short‑form SPD questions. The SPD\_RADJ variable was created by adjusting the long-form responses to match the short-form responses, thus creating a variable that could be combined with the short-form only variable from 2005 (i.e., SPD\_UADJ). [↑](#footnote-ref-13)
14. The associated variable name for SUITHKYR on the analytic file is MHSUITHK. [↑](#footnote-ref-14)
15. The associated variable name for U\_ALCMON on the analytic file is ALCMON. [↑](#footnote-ref-15)
16. The associated variable name for U\_BNGALC on the analytic file is BINGEDRK. [↑](#footnote-ref-16)
17. FIPS codes also have been assigned to outlying areas of the United States, freely associated states, and individual minor outlying island territories (see <https://www.census.gov/geo/reference/ansi_statetables.html>). These geographic areas are *not* included in NSDUH's data collection efforts nor in its datasets. [↑](#footnote-ref-17)
18. For 2010-2011, two sets of state-level small area estimates were produced. The 2010-2011 "published estimates" are available at <https://www.samhsa.gov/data/> and were based on predictors and weights that used the 2000 census as its base. The 2010-2011 "updated estimates" were developed using new predictors and new weights obtained from the 2010 census. [↑](#footnote-ref-18)
19. Correlations were available for 2002-2003 versus 2009-2010 NSDUH data that included falsified cases, as well as 2002-2003 versus 2009-2010 data excluding falsified cases. For more details on falsification and the revision of 2006 to 2010 data, see Section A.7 of the "2010-2011 National Survey on Drug Use and Health: Guide to State Tables and Summary of Small Area Estimation Methodology" at <https://www.samhsa.gov/data/>. Note that the correlations for 2002-2003 versus 2007-2008 and 2002-2003 versus 2008-2009 were also based on falsified data. [↑](#footnote-ref-19)
20. See Chapter 2 for a description of the benchmarked small area estimate (BSAE) variable and all of the other variables parenthetically noted in this appendix. [↑](#footnote-ref-20)
21. Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics). New York, NY: John Wiley & Sons. [↑](#footnote-ref-21)
22. See the reference in footnote 21. [↑](#footnote-ref-22)
23. There are 13 overlapping time periods, as follows: (a) 1999-2000 versus 2000-2001, (b) 2002-2003 versus 2003-2004, (c) 2003-2004 versus 2004-2005, (d) 2004-2005 versus 2005-2006, (e) 2005-2006 versus 2006-2007, (f) 2006-2007 versus 2007-2008, (g) 2007-2008 versus 2008-2009, (h) 2008-2009 versus 2009-2010, (i) 2009-2010 versus 2010-2011, (j) 2010-2011 versus 2011-2012, (k) 2011-2012 versus 2012-2013, (l) 2012-2013 versus 2013-2014, and (m) 2013-2014 versus 2014-2015. [↑](#footnote-ref-23)