National Survey on Drug Use and Health
Generalized Correlations of
Small Area State Estimates between
Nonoverlapping Time Periods:
Documentation for CSV and Excel Files

This page intentionally left blank

Documentation for CSV and Excel Files

Description of the CSV File Type

Files with a comma separated value (*.csv) extension are in plain text. They contain characters stored in a flat, nonproprietary format and can be opened by most computer programs. Each *.csv file contains a set of tabular data, with each record delineated by a line break and each field within a record delineated by a comma. A field that contains commas as part of its content has the additional delineation of a quote mark character before and after the field's contents. When a quote mark character is part of a field's content, it is included as two consecutive ""quote mark"" characters.

Computers with Microsoft Excel installed open *.csv files in Excel by default, with the fields automatically arranged appropriately in columns. Other database programs also open *.csv files with the fields appropriately arranged.

This zip archive holds 26 CSV files (i.e., "NSDUHsaeGenCorrTab#-2015.csv"), reflecting the 26 "Generalized Correlation Table #" tabs in the Excel file, and they contain the table title, table notes, column headings, and data.

Description of Generalized Correlations Used to Compare State Small Area Estimates between Two Nonoverlapping Time Periods

Starting with the comparison of 2002-2003 and 2003-2004 state estimates from the National Survey on Drug Use and Health (NSDUH), tests of significance of the difference in point estimates containing an overlapping year have been produced annually. In addition to these overlapping year comparisons, some nonoverlapping state comparisons with respect to the baseline period 2002-2003 (e.g., 2002-2003 vs. 2007-2008, 2002-2003 vs. 2008-2009, and beyond) have also been produced and are available for downloading from Substance Abuse and Mental Health Services Administration (SAMHSA) at http://www.samhsa.gov/data/.[1] However, users of NSDUH estimates based on small area estimation (SAE) might be interested in conducting tests of significance not published for other nonoverlapping time periods, such as 2006-2007 versus 2009-2010. In order to produce the appropriate test statistic necessary to determine if the difference is statistically significant (e.g., the *p* value), the estimates, the Bayesian confidence interval (CI) for each estimate, and the correlation between the two estimates are needed. The estimates and CIs are available at http://www.samhsa.gov/data/; however, the correlations were not available prior to the release of the 2014-2015 state estimates. These correlations represented by generalized correlations, along with the published small area estimates and Bayesian CIs, should be used to compare state prevalence rates between two

[1] Because of methodological changes implemented in the 2002 survey, a new baseline for all outcomes began that year. For the mental health outcomes, including any mental illness (AMI), serious mental illness (SMI), and serious thoughts of suicide, the baseline is 2008 because of new questions that were introduced in the survey that year.

nonoverlapping time periods. The methodology for conducting such comparisons is illustrated by an example given later in this document.

The correlation in state estimates over time periods results from simultaneously modeling the data associated with the time periods of interest and/or the commonality of the data between the two time periods.[2] The correlation due to this simultaneous modeling results mostly from the random effects for the population subgroups (age group by time period) being correlated over areas. For this simultaneous modeling, four age groups (12 to 17, 18 to 25, 26 to 34, and 35 or older), or three age groups (18 to 25, 26 to 34, and 35 or older) for the mental health outcomes, by two nonoverlapping time periods (i.e., eight or six subpopulation-specific models) were simultaneously 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 or 6×6 matrices corresponding to the eight element or six element vectors of random effects. This correlation indicates that the area-level random contributions to the intercepts for the population subgroup-specific models can still be correlated for nonoverlapping years due to the random intercept adjustments having similar up and down patterns over areas for the two nonoverlapping time periods. Having a fixed common set of predictors across time in the SAE models might contribute to this correlation; however, no commonality of the fixed-effect predictors is required for these population subgroup-specific intercept adjustments to be correlated across areas for nonoverlapping years.

The correlation in state estimates across *overlapping time periods* is a result of simultaneously modeling the data associated with the time periods of interest and the commonality of data associated with the middle year (e.g., in the 2006-2007 vs. 2007-2008 state change estimates, the data for 2007 are common to both sets of estimates). Conversely, the correlation in state estimates across *nonoverlapping time periods* results solely from simultaneously modeling the data associated with the time periods of interest. The overlapping year correlations tend to be larger than the nonoverlapping year correlations because of commonality of the data associated with the middle year. The variance of the difference between state estimates depends on the underlying correlation between the state estimates. If the state estimates are assumed to be noncorrelated or the correlation between the state estimates is assumed to be smaller than the actual correlation, then the difference would likely be declared nonsignificant. In order to obtain reasonable estimates of this difference over nonoverlapping time periods, it is desirable to include appropriate correlations in the estimation methodology, which would require simultaneous modeling of data associated with the time periods of interest. As mentioned earlier, due to budget and time constraints, it is not practical to simultaneously model the data corresponding to all possible combinations of nonoverlapping time periods in advance. As a proxy, because nonoverlapping year correlations are expected to be between the "long-term" change correlations (i.e., correlations between the baseline period of 2002-2003 and a time period several years beyond) and the overlapping year correlations, a conservative estimate of nonoverlapping time period correlations could be the average of the long-term change correlations.

[2] For more information on this type of correlation, see the "National Survey on Drug Use and Health: Comparison of 2002-2003 and 2010-2011 Model-Based Prevalence Estimates (50 States and the District of Columbia)" and the "National Survey on Drug Use and Health: Comparison of 2009-2010 and 2010-2011 Model-Based Prevalence Estimates (50 States and the District of Columbia)" at http://www.samhsa.gov/data/.

Currently, seven sets of long-term change correlations are available for each substance use measure arranged according to outcome by state by age group: (a) 2002-2003 versus 2007-2008, (b) 2002-2003 versus 2008-2009, (c) two sets of 2002-2003 versus 2009-2010,[3] (d) 2002-2003 versus 2010-2011, (e) 2002-2003 versus 2012-2013, and (f) 2002-2003 versus 2013-2014. Correlations for the four mental health outcomes are available for a different set of time periods, as discussed in the next paragraph. The average of these seven sets of correlations is henceforth referred to as a "generalized correlation." Averaging seven sets of correlations minimizes variation and reduces the risk of using an outlier from a particular set of pair-years. Each of these seven 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), that is, 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.

For three of the four mental health measures (i.e., AMI, SMI, and suicidal thoughts), six sets of correlations are available and are arranged according to outcome by state by age group: (a) 2008-2009 versus 2010-2011, (b) 2008-2009 versus 2011-2012, (c) 2008-2009 versus 2012-2013, (d) 2009-2010 versus 2011-2012, (e) 2009-2010 versus 2012-2013, and (f) 2010-2011 versus 2012-2013. The average of these six sets of correlations is the "generalized correlation." Similarly, the fourth mental health measure—major depressive episode (MDE)—has eight sets of correlations available that are arranged by state and age group: (a) 2005-2006 versus 2007-2008, (b) 2005-2006 versus 2008-2009, (c) 2005-2006 versus 2009-2010, (d) 2005-2006 versus 2010-2011, (e) 2005-2006 versus 2011-2012, (f) 2005-2006 versus 2012-2013, (g) 2006-2007 versus 2009-2010, and (h) 2008-2009 versus 2010-2011. The average of these eight sets of correlations is the "generalized correlation." Note that these correlations were produced in the same manner as discussed in the previous paragraph.

These generalized correlations should be used by NSDUH data users to test the null hypothesis of no difference in state (or census region) prevalence rates for any two nonoverlapping time periods (e.g., 2006-2007 vs. 2010-2011). The national estimates are direct estimates, so the correlations for these are zero. To reiterate, these generalized correlations are not to be used for conducting tests of significance between two overlapping time periods (i.e., 2010-2011 vs. 2011-2012).

[3] During regular data collection and processing checks for the 2011 NSDUH, data errors were identified. These errors affected the data for Pennsylvania (2006 to 2010) and Maryland (2008 and 2009) (for more details about the data errors, see Section A.7 of the "2011-2012 National Survey on Drug Use and Health: Guide to State Tables and Summary of Small Area Estimation Methodology" at http://www.samhsa.gov/data/). The first set of 2002-2003 versus 2009-2010 correlations that was produced before the data errors were identified included the erroneous data from Pennsylvania and Maryland. The second set of 2002-2003 versus 2009-2010 correlations was produced excluding the erroneous data from Pennsylvania and Maryland. The two sets of 2002-2003 versus 2009-2010 correlations were compared, and it was concluded that the data errors did not affect the underlying correlations. Therefore, the previously produced correlations (2002-2003 vs. 2007-2008 and 2002-2003 vs. 2008-2009) were not revised.

The methodology that is used to compare state prevalence rates for two time periods is given in the "National Survey on Drug Use and Health: Comparison of 2002-2003 and 2011-2012 Model-Based Prevalence Estimates (50 States and the District of Columbia)" at http://www.samhsa.gov/data/. Note that a different set of generalized correlations was used to produce the *p* values for comparing the 2002-2003 and 2011-2012 small area estimates. Those generalized correlations were an average of five sets of correlations (all sets except the 2002-2003 vs. 2012-2013 correlations were available at the time). Using the methodology provided in that document, NSDUH data users can compare state prevalence rates for any two nonoverlapping time periods. To illustrate the procedure, an example comparing the 2006-2007 and 2011-2012 state prevalence rates of past month illicit drug use in Alabama among young adults aged 18 to 25 is given in the next section. Note that there were changes to the survey in 2002;[4] thus, these correlations should be used to compare state prevalence rates only from 2002-2003 and beyond.

Comparison of State Estimates in Nonoverlapping Years Using a Generalized Correlation

This section describes a method for determining whether differences in prevalence rates 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 $\pi_{sa(1)}$ and $\pi_{sa(2)}$ denote the prevalence rates at time period 1 and time period 2, respectively, for state-*s* and age group-*a*. The difference between $\pi_{sa(1)}$ and $\pi_{sa(2)}$ is defined in terms of the log-odds ratio (lor_{sa}) as opposed to the simple difference because the posterior distribution of lor_{sa} is closer to Gaussian than the posterior distribution of the simple difference $(\pi_{sa(2)} - \pi_{sa(1)})$. The lor_{sa} is defined as

$$lor_{sa} = \ln\left[\frac{\pi_{sa(2)} / (1 - \pi_{sa(2)})}{\pi_{sa(1)} / (1 - \pi_{sa(1)})}\right],$$

where ln denotes the natural logarithm. The *p* value is computed to test the null hypothesis of no change (i.e., $\pi_{sa(2)} = \pi_{sa(1)}$ or equivalently, $lor_{sa} = 0$). An estimate of lor_{sa} is given by

$$\hat{l}or_{sa} = \ln\left[\frac{p_{sa(2)} / (1 - p_{sa(2)})}{p_{sa(1)} / (1 - p_{sa(1)})}\right],$$

[4] For details, see Section A.2 of the "2011-2012 National Surveys on Drug Use and Health: Guide to State Tables and Summary of Small Area Estimation Methodology" at http://www.samhsa.gov/data/.

where $p_{sa(1)}$ and $p_{sa(2)}$ are the state estimates (i.e., the benchmarked small area estimates [BSAEs]) for the 2 years being compared. To compute the variance of $\hat{lor}_{sa}$, that is, $v(\hat{lor}_{sa})$, let $\hat{\theta}_1 = \frac{p_{sa(1)}}{1 - p_{sa(1)}}$ and $\hat{\theta}_2 = \frac{p_{sa(2)}}{1 - p_{sa(2)}}$, then

$$v(\hat{lor}_{sa}) = v[\ln(\hat{\theta}_1)] + v[\ln(\hat{\theta}_2)] - 2\,\mathrm{cov}[\ln(\hat{\theta}_1), \ln(\hat{\theta}_2)],$$

where $\mathrm{cov}[\ln(\hat{\theta}_1), \ln(\hat{\theta}_2)]$ denotes the covariance between $\ln(\hat{\theta}_1)$ and $\ln(\hat{\theta}_2)$. This covariance is defined in terms of the associated correlation as follows:

$$\mathrm{cov}[\ln(\hat{\theta}_1), \ln(\hat{\theta}_2)] = \mathrm{correlation}\,[\ln(\hat{\theta}_1), \ln(\hat{\theta}_2)] \times \sqrt{v[\ln(\hat{\theta}_1)] \times v[\ln(\hat{\theta}_2)]},$$

where $v[\ln(\hat{\theta}_i)] = \left(\frac{U_i - L_i}{2 \times 1.96}\right)^2$ for $i = 1, 2$, $U_i = \ln \frac{upper_i}{1 - upper_i}$, $L_i = \ln \frac{lower_i}{1 - lower_i}$, and the *lower* and *upper* are the 95 percent Bayesian CIs.

For the correlation between $\ln(\hat{\theta}_1)$ and $\ln(\hat{\theta}_2)$ for an outcome measure by state by age group, the generalized correlation will be used.

To calculate the *p* value for testing the null hypothesis of no difference $(lor = 0)$, it is assumed that the posterior distribution of lor is normal with $mean = \hat{lor}_{sa}$ and $variance = v(\hat{lor}_{sa})$. With the null value of $(lor = 0)$, the Bayes *p* value or significance levels for the null hypothesis of no difference is $p\ value = 2 * P[Z \geq abs(z)]$, where Z is a standard normal random variate, $z = \frac{\hat{lor}_{sa}}{\sqrt{v(\hat{lor}_{sa})}}$, and $abs(z)$ denotes the absolute value of z. This Bayesian significance level (or *p* value) for the null value of lor, say lor_0, is defined following Rubin[5] as the posterior probability for the collection of the lor values that are less likely or have smaller posterior density $d(lor)$ than the null (no change) value lor_0. That is, $p\,value(lor_0) = probability[d(lor) \leq d(lor_0)]$. With the posterior distribution of lor approximately normal, $p\ value(lor_0)$ is given by the above expression.

[5] 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.

For overlapping time periods,[6] 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.

Example. The following exhibit shows the prevalence estimates for past month illicit drug use among young adults aged 18 to 25 in Alabama for 2006-2007 and 2011-2012.

State	Estimate (%)	95% Confidence Interval (%)
2006-2007[1]	15.90	(13.18, 19.05)
2011-2012[2]	17.51	(14.96, 20.40)

[1] See Table 1 of the "2006-2007 NSDUH: Model-Based Prevalence Estimates (50 States and the District of Columbia" at http://www.samhsa.gov/data.

[2] See Table 1 of the "2011-2012 NSDUH: Model-Based Prevalence Estimates (50 States and the District of Columbia)" at http://www.samhsa.gov/data.

The generalized correlation for illicit drug use for 18 to 25 years olds in Alabama is 0.21994.[7] Note that generalized correlations are on the logit scale;[8] that is, they are the correlation between the logit of p_1 and the logit of p_2 (not the correlation between p_1 and p_2, where p_1 and p_2 are the 2006-2007 and the 2011-2012 small area estimates, respectively).

The p value is calculated using the following methodology. Using the data from the exhibit, the following terms are first defined:

$$p_1 = 0.1590,\ lower_1 = 0.1318,\ upper_1 = 0.1905,\ p_2 = 0.1751,\ lower_2 = 0.1496,\ \text{and}\ upper_2 = 0.2040.$$

Then the following calculations are made:

$$\hat{lor} = \ln\left[\frac{p_2/(1-p_2)}{p_1/(1-p_1)}\right] = \ln\left[\frac{0.1751/(1-0.1751)}{0.1590/(1-0.1590)}\right] = 0.1158,$$

[6] The overlapping time periods are as follows: 1999-2000 versus 2000-2001, 2002-2003 versus 2003-2004, 2003-2004 versus 2004-2005, 2004-2005 versus 2005-2006, 2005-2006 versus 2006-2007, 2006-2007 versus 2007-2008, 2007-2008 versus 2008-2009, 2008-2009 versus 2009-2010, 2009-2010 versus 2010-2011, 2010-2011 versus 2011-2012, 2011-2012 versus 2012-2013, and 2012-2013 versus 2013-2014.

[7] See Table 1 of the generalized correlation Excel files at http://www.samhsa.gov/data.

[8] The logit scale is defined as follows: $\text{logit}(x) = \ln(\frac{x}{1-x})$, where ln denotes the natural logarithmic function.

$$U_1 = \ln\frac{upper_1}{1-upper_1} = \ln\frac{0.1905}{1-0.1905} = \text{-1.44676},$$

$$L_1 = \ln\frac{lower_1}{1-lower_1} = \ln\frac{0.1318}{1-0.1318} = \text{-1.88514},$$

$$U_2 = \ln\frac{upper_2}{1-upper_2} = \ln\frac{0.2040}{1-0.2040} = \text{-1.36148, and}$$

$$L_2 = \ln\frac{lower_2}{1-lower_2} = \ln\frac{0.1496}{1-0.1496} = \text{-1.73774.}$$

Define $\hat{\theta}_1 = p_1/(1-p_1)$ and $\hat{\theta}_2 = p_2/(1-p_2)$, then the variance of $\hat{\theta}_1$ and $\hat{\theta}_2$ is given by the following:

$$v[\ln(\hat{\theta}_1)] = \left(\frac{U_1 - L_1}{2\times 1.96}\right)^2 = \left(\frac{-1.44676+1.88514}{2\times 1.96}\right)^2 = 0.01251 \text{ and}$$

$$v[\ln(\hat{\theta}_2)] = \left(\frac{U_2 - L_2}{2\times 1.96}\right)^2 = \left(\frac{-1.36148+1.73774}{2\times 1.96}\right)^2 = 0.00921.$$

Using the above variances and the generalized correlation, the variance of $\hat{lor}$ is given by the following:

$$v(\hat{lor}) = v[\ln(\hat{\theta}_1)] + v[\ln(\hat{\theta}_2)] - 2\,\mathrm{cov}[\ln(\hat{\theta}_1), \ln(\hat{\theta}_2)],$$

where

$$\begin{aligned}\mathrm{cov}[\ln(\hat{\theta}_1), \ln(\hat{\theta}_2)] &= correlation[\ln(\hat{\theta}_1), \ln(\hat{\theta}_2)] \times \sqrt{v[\ln(\hat{\theta}_1)] \times v[\ln(\hat{\theta}_2)]} \\ &= 0.21994 \times \sqrt{0.01251 \times 0.00921} = 0.00236.\end{aligned}$$

Hence,

$$v(\hat{lor}) = 0.01251 + 0.00921 - 2\times 0.00236 = 0.01700 \text{ and}$$

$$z = \frac{\hat{lor}}{\sqrt{v[\hat{lor}]}} = \frac{0.1158}{\sqrt{0.01700}} = 0.88808.$$

The Bayes *p* value for the null hypothesis of no difference is defined as follows: $p\ value = 2\times P[Z \geq abs(0.88808)] = 0.375$, where *abs* denotes the absolute value and *Z* is the standard normal random variable. Because the *p* value is greater than 0.05, it can be said that at the 5 percent level of significance, these two prevalence rates are not significantly different.

This page intentionally left blank