

# 2004 NATIONAL SURVEY ON DRUG USE AND HEALTH

## STATISTICAL INFERENCE REPORT

Prepared for the 2004 Methodological Resource Book

Contract No. 283-03-9028  
RTI Project No. 0208726.181.003  
Deliverable No. 26

Authors:

Walter Boyle  
James R. Chromy  
Andrew Clarke  
Teresa R. Davis  
Steven L. Emrich  
Erica Hirsch  
Susan Hunter  
Breda Munoz  
Lisa E. Packer

Project Director: Thomas G. Virag

Prepared for:

Substance Abuse and Mental Health Services Administration  
Rockville, MD 20857

Prepared by:

RTI International  
Research Triangle Park, NC 27709

December 2005

# 2004 NATIONAL SURVEY ON DRUG USE AND HEALTH

## STATISTICAL INFERENCE REPORT

RTI Project No. 0208726.181.003  
Contract No. 283-03-9028

Deliverable No. 26

**Authors:**

Walter Boyle  
James R. Chromy  
Andrew Clarke  
Teresa R. Davis  
Steven L. Emrich  
Erica Hirsch  
Susan Hunter  
Breda Munoz  
Lisa E. Packer

**Project Director:**

Thomas G. Virag

Prepared for:

Substance Abuse and Mental Health Services Administration  
Rockville, MD 20857

Prepared by:

RTI International  
Research Triangle Park, NC 27709

December 2005

## **Acknowledgments**

This publication was developed for the Substance Abuse and Mental Health Services Administration (SAMHSA), Office of Applied Studies (OAS), by RTI International (a trade name of Research Triangle Institute), Research Triangle Park, North Carolina, under Contract No. 283-03-9028. Significant contributors at RTI include Walter Boyle, James R. Chromy, Andrew Clark, Teresa R. Davis, Steven L. Emrich, Erica Hirsch, Susan Hunter, Breda Munoz, and Lisa E. Packer. Joyce Clay-Brooks word processed and formatted the report, Richard S. Straw copyedited the report, Diane Eckard worked on the graphic, and Thomas G. Virag is the RTI Project Director.

## Table of Contents

Section	Page
1. Introduction.....	1
2. Background.....	3
3. Prevalence Rates .....	5
4. Sampling Error.....	7
5. Degrees of Freedom.....	9
6. Statistical Significance of Differences.....	11
7. Confidence Intervals .....	13
8. Calendar Year and Past Year Incidence Estimates .....	15
9. Suppression of Estimates with Low Precision.....	19
References.....	23

# 1. Introduction

Statistical inference occurs whenever data obtained from sample observations belonging to and considered representative of a larger target population are used to make generalizations concerning the larger population. The target population for the 2004 National Survey on Drug Use and Health (NSDUH)<sup>1</sup> was the U.S. civilian, noninstitutionalized population aged 12 or older (at the time of their interview) in 2004. Measurements for this target population were the responses to the survey questions provided by persons participating in the 2004 survey.

Statistical inferences concerning characteristics of interest for this population and various subpopulations are presented in the form of estimates derived from the sample data collected. Examples of the inferences made from the 2004 NSDUH data include estimates of the number of persons who were substance users during the past month, past year, and their lifetime, and the associated percentages (prevalence rates) of substance use for these reference periods. Inferences also were made for such categories as substance initiation; risk and protective factors; substance dependence, abuse, and treatment; and measures related to mental health problems. Among some populations of interest, sample sizes were not adequate to support inferences; in these cases, estimates were produced from annual averages based on combined data.

This report is organized as follows. Section 2 provides background information concerning the 2004 NSDUH; Sections 3 and 4 discuss the prevalence rates and sampling errors and how they were calculated; Section 5 describes the degrees of freedom that were used when comparing estimates; and Section 6 discusses how statistical significance of differences between estimates was determined. Section 7 discusses confidence interval estimation, and Section 8 describes how the rates for initiation or incidence of drug use were computed. Finally, Section 9 discusses the conditions under which estimates with low precision were suppressed.

---

<sup>1</sup> Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA).



## 2. Background

The 2004 National Survey on Drug Use and Health (NSDUH) was a continuation of a coordinated 5-year, 50-State sample design to provide national and State estimates of drug use for the survey years from 1999 through 2003.<sup>2</sup> Additionally, the use of computer-assisted interviewing (CAI) methods for the screening and interviewing of selected respondents was continued in an effort to maintain consistency and preserve trend analyses between the 2004 NSDUH and earlier survey years.

For the 50-State design, eight States were designated as large sample States: California, Florida, Illinois, Michigan, New York, Ohio, Pennsylvania, and Texas. The samples collected from each of these States were large enough to support direct State estimates. In 2004, sample sizes in the eight large States ranged from 3,575 to 3,725. For the remaining 42 States and the District of Columbia, smaller samples were selected, but these were sufficient to support State estimates using small area estimation (SAE) techniques. Sample sizes in these small States ranged from 828 to 934 in 2004.

In the 50-State design, the geographic area as a whole was first stratified into a total of 900 field interviewer (FI) regions (48 regions in each large sample State and 12 regions in each small sample State). Within FI regions, adjacent census blocks were combined to form the first-stage sampling units called "segments." Eight sample segments per FI region were fielded during the 2004 survey year. These sampled segments were allocated equally into four separate samples, one for each 3-month period during the year, so that the survey remained in the field in each FI region year-round.

Although most of the methods and techniques of the 2003 NSDUH were continued for the 2004 NSDUH (e.g., the use of a \$30 respondent incentive<sup>3</sup>), several changes in the sample design were implemented. Specifically, the sample design included an additional step in which approximately 50 percent of the adult respondents aged 18 or older were randomly assigned to receive the full module of questions on serious psychological distress (SPD).<sup>4</sup> The remaining adults received a reduced number of SPD questions and a new set of questions on depression. These complementary samples are together referred to as the SPD "split sample," the full SPD module is referred to as "sample A," and the reduced SPD module is referred to as "sample B."<sup>5</sup>

The final respondent sample of 67,760 persons for the 2004 NSDUH provides a sufficient sample to create domain estimates for a broad range of ages and other demographic categories. Individual observations are weighted in a manner such that the weighted sample is representative of the civilian, noninstitutionalized population aged 12 or older for both the general U.S.

---

<sup>2</sup> For more information on the sample design in the 2004 NSDUH, see the sample design report in the *2004 National Survey on Drug Use and Health: Methodological Resource Book* (Bowman, Chromy, Hunter, & Martin, 2005), which is available at <http://www.oas.samhsa.gov/nhsda/methods.cfm#2k4>.

<sup>3</sup> For more information on the implications and use of respondent incentives in NSDUH, see the *2001 National Household Survey on Drug Abuse: Incentive Experiment Combined Quarter 1 and Quarter 2 Analysis* (Office of Applied Studies [OAS], 2002).

<sup>4</sup> The 2004 CAI originally referred to the serious psychological distress module as serious mental illness.

<sup>5</sup> For details, see the report mentioned in footnote 2.

population as well as for each of the individual States.<sup>6</sup> However, for certain populations of interest, 2 years of NSDUH data were combined to obtain annual averages. The person-level weights for estimates based on the annual averages were obtained by dividing the analysis weights for each of the 2 specific years by a factor of two.

Due to the new split sample described above, two additional sets of analysis weights were required to create domain estimates for SPD and the adult depression module (i.e., major depressive episodes, or MDE). The weights for sample A were used as the analysis weights for producing the SPD estimates, and the weights for sample B were used as the analysis weights for producing the MDE estimates. These two weights were created by incorporating the inverse quarterly sampling fractions associated with the split samples for the two modules into the design weights after the person-level nonresponse adjustment. Each subsample then was poststratified separately to the census estimates of the civilian noninstitutionalized population aged 18 or older for various domains defined by age group, race/ethnicity, gender, and State. Note that there were six respondents aged 18 or older who had a missing value for the split-sample indicator variable. It appears that these six respondents broke off the interview before they could be assigned to the full or reduced SPD module. Those six respondents were excluded from either sample A or sample B; thus, they had zero weight of sample A or sample B.

---

<sup>6</sup> For more information on the sampling weight calibration in the 2004 NSDUH, see the person-level sampling weight calibration report in the *2004 National Survey on Drug Use and Health: Methodological Resource Book* (Chen et al., 2005), which is available at <http://www.oas.samhsa.gov/nhsda/methods.cfm#2k4>.

### 3. Prevalence Rates

The national prevalence rates were computed using a multiprocedure package called SURvey DATA ANalysis (SUDAAN<sup>®</sup>) Software for Statistical Analysis of Correlated Data (RTI International, 2004b). The final, nonresponse-adjusted, and poststratified analysis weights were used in SUDAAN to compute unbiased design-based drug use estimates.

Prevalence rates are the proportions of the population who exhibit characteristics of interest (such as substance use). Let  $\hat{p}_d$  represent the prevalence rate of interest for domain  $d$ . Then  $\hat{p}_d$  would be defined as the ratio

$$\hat{p}_d = \frac{\hat{Y}_d}{\hat{N}_d},$$

where  $\hat{Y}_d$  = estimated number of persons exhibiting the characteristic of interest in domain  $d$ , and  $\hat{N}_d$  = estimated population total for domain  $d$ .

$\hat{N}_d$  is estimated as  $\sum w_i \delta_i$ , where  $w_i$  represents the analysis weight and  $\delta_i$  represents an indicator variable, which is defined as

$$\delta_i(d) = \begin{array}{l} 1 \text{ if the } i^{\text{th}} \text{ sample unit is in subgroup } d, \\ 0 \text{ otherwise.} \end{array}$$

For certain populations of interest, the 2002-2003 and 2003-2004 data were combined to obtain annual averages, then the prevalence rates were computed in SUDAAN as described above. The annual averages were derived by concatenating the 2002-2003 and 2003-2004 datasets and then dividing the analysis weights by a factor of 2.



## 4. Sampling Error

As were the prevalence rates, all of the variance estimates (including those for prevalence based on annual averages from combined data) were calculated using a method in SUDAAN that is unbiased for linear statistics. This method is based on multistage clustered sample designs where the first-stage (primary) sampling units are drawn with replacement.

Due to the complex nature of the sampling design for the National Survey on Drug Use and Health (NSDUH) (specifically the use of stratified-clustering sampling), key nesting variables were created for use in SUDAAN to capture explicit stratification and to identify clustering. For the 2004 NSDUH, each field interviewer (FI) region was considered its own stratum with two replicates defined within each variance stratum (FI region). The first replicate consisted of those "phasing-out" segments (i.e., those that would not be used in the next survey year). The second replicate was made up of those "phasing-in" segments (i.e., those that would be fielded again the following year), thus constituting the 50 percent overlap between survey years. Each variance replicate consisted of four segments, one segment for each quarter of data collection.

Estimates of means or proportions,  $\hat{p}_d$ , such as drug use prevalence rates, take the form of nonlinear statistics whenever the variances cannot be expressed in closed form. Variance estimation for nonlinear statistics in SUDAAN is performed using a first-order Taylor series approximation of the deviations of estimates from their expected values (RTI International, 2004b).

Estimates of domain totals,  $\hat{Y}_d$ , corresponding to estimates of domain means or proportions,  $\hat{p}_d$ , can be estimated as

$$\hat{Y}_d = \hat{N}_d \cdot \hat{p}_d,$$

where  $\hat{N}_d$  = estimated population total for domain  $d$ , and  $\hat{p}_d$  = estimated mean or proportion for domain  $d$ .

The standard error (SE) for the total estimate is obtained by multiplying the SE of the mean or proportion by  $\hat{N}_d$ , that is,

$$SE(\hat{Y}_d) = \hat{N}_d \cdot SE(\hat{p}_d).$$

This approach is theoretically correct when the domain size estimates,  $\hat{N}_d$ , are among those forced to Census Bureau population projections through the weight calibration process.<sup>7</sup> In these cases,  $\hat{N}_d$  is not subject to sampling error.

For estimated domain totals,  $\hat{Y}_d$ , where  $\hat{N}_d$  is not fixed (i.e., where domain size estimates are not forced to Census Bureau population projections), this formula may still provide a good approximation if it can be assumed that the sampling variation in  $\hat{N}_d$  is negligible relative to the sampling variation in  $\hat{p}_d$ . For most NSDUH estimates, this is a reasonable assumption.

However, for a subset of tables produced from the 2004 data, the above approach yielded an underestimate of the variance of a total because  $\hat{N}_d$  was subject to considerable variation. In these cases, a direct estimate of the SE of  $\hat{Y}_d$  was taken from SUDAAN.

In previous years, all SEs of estimates produced for a particular table were calculated in the same manner (either as the Taylor series approximation directly from SUDAAN or the product of the mean or proportion and  $\hat{N}_d$ ); however, in 2004 it was determined that a number of estimates created for populations controlled for in the weighting process required further consideration. Specifically, these estimates should have corresponding SEs of zero, but due to the methods of calculation, nonzero values for these SEs were being reported.

Because this issue was identified during the 2004 table production, the Substance Abuse and Mental Health Services Administration (SAMHSA) decided to implement a change in the method of SE calculation on an estimate-by-estimate basis for the 19 sample size and population tables found in Section 8 of the 2004 detailed tables<sup>8</sup> and for any new tables in 2004. Specifically, this change was designed to indicate when estimated population sizes (corresponding to domain estimates that have been forced to Census Bureau population projections) result in SEs that are zero. Tables not included in Section 8 of the 2004 detailed tables that were created in prior years were not adjusted, but will be given additional consideration in 2005.

---

<sup>7</sup> For more information on the sampling weight calibration in the 2004 NSDUH, see the person-level sampling weight calibration report in the *2004 National Survey on Drug Use and Health: Methodological Resource Book* (Chen et al., 2005), which is available at <http://www.oas.samhsa.gov/nhsda/methods.cfm#2k4>.

<sup>8</sup> The 2004 NSDUH's detailed tables are available on the SAMHSA website at <http://www.oas.samhsa.gov/WebOnly.htm#NHSDAtabs>. Detailed tables for the 1998 to 2003 surveys are available at the same location.

## 5. Degrees of Freedom

To determine whether the observed difference between estimates is statistically significant, the degrees of freedom (*df*) are needed to locate the corresponding probability level (*p* value) of the test statistic. The test statistic is computed from the sample data and represents a numerical summary of the difference between the estimates under consideration; it is a random variable that has a predetermined distribution (such as student's *t*, chi-square, or *F*). The degrees of freedom characterize the amount of variation expected in the estimation of sampling error and are used in conjunction with the test statistic to determine probabilities and evaluate statistical significance.

The degrees of freedom are calculated as the number of primary sampling units (PSUs, variance replicates) minus the number of strata for the data being analyzed. In National Survey on Drug Use and Health (NSDUH) analyses, the degrees of freedom are based on the first-level stratification (i.e., the field interviewer [FI] regions). When producing NSDUH estimates on the national level, including estimates based on annual averages from combined data, there are 900 degrees of freedom. If an analysis only involves certain States, the degrees of freedom change depending on whether the State is a large sample or small sample State. The large sample States (i.e., California, Florida, Illinois, Michigan, New York, Ohio, Pennsylvania, and Texas) each have 48 degrees of freedom. All of the other States (i.e., the small sample States, which include the District of Columbia) have 12 degrees of freedom. For specific State analyses (or other subpopulations of interest), the degrees of freedom can be specifically indicated in SUDAAN; otherwise, the degrees of freedom are computed using the entire dataset.



## 6. Statistical Significance of Differences

Once the degrees of freedom have been determined, various methods used to compare prevalence estimates may be employed. This section describes some of these methods. Customarily, the observed difference between estimates is evaluated in terms of its statistical significance. Statistical significance is based on the  $p$  value of the test statistic and refers to the probability that a difference as large as that observed would occur due to random variability in the estimates if there were no difference in the prevalence rates being compared. The significance of observed differences is generally reported at the .05 and .01 levels.

Significance tests were conducted on differences between prevalence estimates from the 2003 and 2004 National Surveys on Drug Use and Health (NSDUHs); in a small number of incidence tables, significance tests were conducted on differences between prevalence estimates from 2002 and 2003 data. Significance tests also were conducted on prevalence estimates based on annual averages derived from the combined 2002-2003 and 2003-2004 survey data. Within-year tests were conducted on differences between prevalence estimates for various populations (or subgroups) of interest using data from the 2004 survey. Due to survey design changes implemented in 2002, data from the 2002, 2003, and 2004 NSDUHs should not be compared with data from earlier survey years.

When comparing prevalence estimates, one can test the null hypothesis (no difference between rates) against the alternative hypothesis (there is a difference in prevalence rates) using the standard  $t$  test (with the appropriate degrees of freedom) for the difference in proportions, expressed as

$$t_{df} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\text{var}(\hat{p}_1) + \text{var}(\hat{p}_2) - 2\text{cov}(\hat{p}_1, \hat{p}_2)}}$$

where  $df$  = the appropriate degrees of freedom,  $\hat{p}_1$  = first prevalence estimate,  $\hat{p}_2$  = second prevalence estimate,  $\text{var}(\hat{p}_1)$  = variance of first prevalence estimate,  $\text{var}(\hat{p}_2)$  = variance of second prevalence estimate, and  $\text{cov}(\hat{p}_1, \hat{p}_2)$  = covariance between  $\hat{p}_1$  and  $\hat{p}_2$ . Note that the first and second prevalence estimates may take the form of prevalence estimates from two different survey years (e.g., 2003 and 2004, respectively), prevalence estimates from sets of combined survey data (e.g., 2002-2003 annual averages and 2003-2004 annual averages, respectively), or prevalence estimates for populations of interest within a single survey year.

Under the null hypothesis,  $t$  is distributed as a random variable from the  $t$ -distribution. Therefore, calculated values of  $t$ , along with the appropriate degrees of freedom, can be used to determine the corresponding probability level (i.e.,  $p$  value). Whether testing for differences between years or from different populations within the same year, the covariance term in the formula for  $t$  will, in general, not be equal to zero. SUDAAN is used to compute estimates of  $t$  along with the associated  $p$  values such that the covariance term is calculated by taking the sample design into account. A similar procedure and formula for  $t$  are used for estimated totals.

As the degrees of freedom approach infinity, the  $t$  distribution approaches the standard normal ( $Z$ ) distribution. That is, because most of the statistical tests performed have 900 degrees of freedom, the  $t$  tests performed produce approximately the same numerical results as if a  $Z$  test had been performed.

When comparing population subgroups defined by three or more levels of a categorical variable, log-linear chi-square tests of independence of the subgroup and the prevalence variables were conducted first to control the error level for multiple comparisons. If a chi-square test indicated overall significant differences, the significance of each particular pairwise comparison of interest was tested using SUDAAN analytic procedures to properly account for the sample design. A detailed description of the test statistic, which is based on the Wald statistic, can be found in the SUDAAN language manual (RTI International, 2004a, p. 177).

## 7. Confidence Intervals

In some National Survey on Drug Use and Health (NSDUH) publications, sampling error has been quantified using 95 percent confidence intervals (CIs). Because NSDUH estimates are frequently small percentages, the confidence intervals are based on logit transformations. Logit transformations yield asymmetric interval boundaries that are more balanced with respect to the probability that the true value falls below or above the interval boundaries than is the case for standard symmetric confidence intervals for small proportions.

To illustrate the method, let the proportion  $P_d$  represent the true prevalence rate for a particular analysis domain  $d$ . Then the logit transformation of  $P_d$ , commonly referred to as the "log odds," is defined as

$$L = \ln[P_d / (1 - P_d)],$$

where "ln" denotes the natural logarithm.

Letting  $\hat{p}_d$  be the estimate of the domain proportion, the log odds estimate becomes

$$\hat{L} = \ln[\hat{p}_d / (1 - \hat{p}_d)].$$

The lower and upper confidence limits of  $L$  are formed as

$$A = \hat{L} - K \left[ \frac{\sqrt{\text{var}(\hat{p}_d)}}{\hat{p}_d(1 - \hat{p}_d)} \right],$$

$$B = \hat{L} + K \left[ \frac{\sqrt{\text{var}(\hat{p}_d)}}{\hat{p}_d(1 - \hat{p}_d)} \right],$$

where  $\text{var}(\hat{p}_d)$  is the variance estimate of  $\hat{p}_d$ , the quantity in brackets is a first-order Taylor series approximation of the standard error (SE) of  $\hat{L}$ , and  $K$  is the constant chosen to yield a level of confidence based on the degrees of freedom ( $df$ ) (e.g.,  $K = 1.96$  for 95 percent confidence limits for national estimates with 900 degrees of freedom).

Applying the inverse logit transformation to  $A$  and  $B$  above yields a confidence interval for  $\hat{p}_d$  as follows:

$$\hat{p}_{d,lower} = \frac{1}{1 + \exp(-A)},$$

$$\hat{p}_{d,upper} = \frac{1}{1 + \exp(-B)},$$

where "exp" denotes the inverse log transformation. The lower and upper confidence interval endpoints for percentage estimates are obtained by multiplying the lower and upper endpoints of  $\hat{p}_d$  by 100.

The confidence interval for the estimated domain total,  $\hat{Y}_d$ , as estimated by

$$\hat{Y}_d = \hat{N}_d \cdot \hat{p}_d,$$

is obtained by multiplying the lower and upper limits of the proportion confidence interval by  $\hat{N}_d$ . For domain totals  $\hat{Y}_d$ , where  $\hat{N}_d$  is not fixed, the confidence interval approximation assumes that the sampling variation in  $\hat{N}_d$  is negligible relative to the sampling variation in  $\hat{p}_d$ .

## 8. Calendar Year and Past Year Incidence Estimates

To assist in the evaluation of trends in the initiation of drug use, National Survey on Drug Use and Health (NSDUH) data also were used to generate estimates of drug use incidence or initiation (i.e., the number of new users during a given year). Incidence rates measure the rapidity with which the numbers of new drug users arise and can suggest emerging patterns of drug use.

The measure of incidence is defined as the number of new cases of drug initiation divided by the person time of exposure. For diseases, the incidence rate,  $IR$ , for a population is defined as the number of new cases of the disease,  $N$ , divided by the person time,  $PT$ , of exposure, or

$$IR = \frac{N}{PT}.$$

The person time of exposure is measured as the net time that individuals in the population during an observed period of time are at risk of developing the disease. This period of time can be for the full period of the study or for a shorter period. The person time of exposure ends at the time of diagnosis (e.g., Greenberg, Daniels, Flanders, Eley, & Boring, 1996, pp. 16-19). Similar conventions were followed for defining the incidence of first use of a substance.

Beginning in 1999, the NSDUH questionnaire allowed for the collection of year and month of first use for recent initiates. The month, day, and year of birth for the initiates also were obtained directly or imputed during the processing of the data. In addition, the questionnaire call record provided the date of the interview. By imputing a day of first use within the year and month of first use reported or imputed, the key respondent inputs, in terms of exact dates, can be computed. Exposure time can be determined in terms of days and converted to an annual value.

Beginning in 2003, the immigrant population was addressed in the incidence analysis. That is, immigrants who initiated drug use outside the United States were not included in the analysis. However, those immigrants who did not initiate outside the United States were included in the analysis for the time period since they entered the United States. If respondents indicated that they were not born in the United States, the survey questionnaire asked them how long they had lived in the United States. Using this information, an imputation-revised entry age and date were created.

Having exact dates of birth and first use (and, if the respondent is an immigrant, his or her exact date of entry into the United States) also allowed the person time of exposure during the targeted period,  $t$ , to be determined. Let the target time period for measuring incidence be specified in terms of dates; for example, the period 1998 would be specified as

$$t = [t_1, t_2) = [1 \text{ Jan } 1998, 1 \text{ Jan } 1999),$$

a period that includes January 1, 1998, and all days up to but not including January 1, 1999. The target age group also can be defined by a half-open interval as  $a = [a_1, a_2)$ . For example, the age group 12 to 17 would be defined by  $a = [12, 18)$  for youths at least age 12, but not yet age 18.

If person  $i$  was in age group  $a$  and residing in the United States during period  $t$ , the time and age interval,  $L_{t,a,i}$ , then can be determined by the intersection

$$L_{t,a,i} = [t_1, t_2) \cap \{ [DOB_i MOB_i YOB_i + a_1, DOB_i MOB_i YOB_i + a_2) \cap [I_{\{US\text{ Born, Immigrant}\}}(i), \infty) \},$$

where the time of birth and time of entry into the United States is defined in terms of day ( $DOB_i$  and  $DOE_i$ ), month ( $MOB_i$  and  $MOE_i$ ), and year ( $YOB_i$  and  $YOE_i$ ), and

$$I_{\{U.S. Born, immigrant\}}(i) = \begin{cases} DOB_i MOB_i YOB_i & \text{if } i \text{ is U.S. Born,} \\ DOE_i MOE_i YOE_i & \text{if } i \text{ is an immigrant.} \end{cases}$$

Either this intersection was empty ( $L_{t,a,i} = \emptyset$ ), or it was defined by the half-open interval,  $L_{t,a,i} = [m_{1,i}, m_{2,i})$ , where

$$m_{1,i} = \text{Max}\{t_1, (DOB_i MOB_i YOB_i + a_1), DOE_i MOE_i YOE_i\},$$

and

$$m_{2,i} = \text{Min}\{t_2, (DOB_i MOB_i YOB_i + a_2)\}.$$

The date of first use,  $t_{fu,d,i}$ , also is expressed as an exact date. An incident of first use of drug  $d$  by person  $i$  in age group  $a$  occurs in time  $t$  if  $t_{fu,d,i} \in [m_{1,i}, m_{2,i})$ . The indicator function,  $I_i(d, a, t)$ , used to count incidents of first use, is set to 1 when  $t_{fu,d,i} \in [m_{1,i}, m_{2,i})$ , and to 0 otherwise. The person time of exposure, measured in years and denoted by  $e_i(d, a, t)$  for a person  $i$  of age group  $a$ , depends on the date of first use of drug  $d$ . If the date of first use precedes the target period ( $t_{fu,d,i} < m_{1,i}$ ), then  $e_i(d, a, t) = 0$ . If the date of first use occurs after the target period or if person  $i$  has never used drug  $d$ , then

$$e_i(d, a, t) = \frac{m_{2,i} - m_{1,i}}{365}.$$

If the date for first use occurs during the target period,  $L_{t,a,i}$ , then

$$e_i(d, a, t) = \frac{t_{fu,d,i} - m_{1,i}}{365}.$$

During leap years, the denominator used to compute person time of exposure is set to 366. Note that both  $I_i(d, a, t)$  and  $e_i(d, a, t)$  are set to 0 if the target period,  $L_{t,a,i}$ , is empty (i.e., person  $i$  is

not in age group  $a$  during time  $t$ ). The incidence rate then is estimated as a weighted ratio estimate:

$$IR(d, a, t) = \frac{\sum_i w_i I_i(d, a, t)}{\sum_i w_i e_i(d, a, t)},$$

where  $w_i$  is the respondent's analytic weight for 2004.

Because the incidence estimates are based on retrospective reports by survey respondents, as was the case for earlier estimates, the estimates may be subject to the same kinds of biases. Differential mortality bias occurs because some persons who were exposed to the risk of first drug use in historical periods died before the 2002 or 2003 NSDUHs were conducted. This type of bias is probably very small. Incidence estimates also are affected by memory errors, including recall decay (tendency to forget events occurring long ago) and forward telescoping (tendency to report that an event occurred more recently than it actually did). Recall decay would tend to result in a downward bias in estimates for earlier years (i.e., 1960s and 1970s), and telescoping would tend to result in an upward bias for estimates in more recent years. There is also likely to be some underreporting bias because of the social stigma of drug use behaviors and respondents' fears of disclosure. This bias is likely to have the greatest impact on recent estimates that reflect more recent use and reporting by younger respondents. Finally, for drug use that is frequently initiated at age 10 or younger, estimates based on 1-year retrospective reports underestimate total incidence because children 11 years old or younger are not sampled by NSDUH. Prior analyses showed that incidence estimates for any alcohol use and any cigarette use could be affected significantly by this. Therefore, for these drugs, only 2003 age-specific rates and the number of initiates aged 18 or older (or 21 or older for applicable tables) were reported.

A recent evaluation of NSDUH retrospective estimates of incidence suggests that these types of bias are significant and differ by substance and length of recall (Gfroerer, Hughes, Chromy, Heller, & Packer, 2004). For very recent time periods, bias in estimates of marijuana, cocaine, alcohol, and cigarette use appear to be small, but for all other types of substance use there is significant downward bias. Bias for all substance use increases the further back in time the estimates are made, suggesting a relationship with the length of recall. Due to the potential reporting biases described above, comparisons between years, particularly between recent estimates and those 10 or more years prior, should be made with caution.

Beginning with the 2004 NSDUH, a new measure related to incidence is being calculated. This measure, termed "past year initiation," refers to respondents whose date of first use of a substance,  $t_{fu,d,i}$ , was within the year prior to their interview. Past year initiation can be viewed as an indicator variable defined as follows:

$$I_{(Past\ Year\ Initiate)}(i) = \begin{cases} 1 & \text{if } (DOI_i MOI_i YOI_i - t_{fu,d,i}) \leq 365 \\ 0 & \text{otherwise} \end{cases},$$

where  $DOI_i$ ,  $MOI_i$ , and  $YOI_i$  denote the day, month, and year of the interview, respectively.

This measure differs from other incidence measures in that it does not refer to a particular calendar year but rather a time period equivalent to the year prior to the interview. Further, these estimates aid in the evaluation of drug use initiation trends and, although calculated in a manner similar to the calendar year incidence estimates, the denominator used is the number of persons in a particular population as opposed to the person time exposure contributed by members of a particular population. As a result, these estimates are calculated using the same methodology as other estimated prevalence rates, discussed in detail in Section 3 on prevalence rates.

One additional difference to be noted is that the calculation of past year initiation does not take into account whether the respondent initiated substance use while a resident of the United States. This has little effect on past year estimates and provides direct comparability with other standard measures of substance use because the populations of interest for the measures will be the same (i.e., both measures examine all possible respondents and do not restrict to those only initiating substance use in the United States). Further, because estimates of past year initiation also are based on retrospective reports of age at first drug use by survey respondents, they may be subject to the same memory-related and/or underreporting biases described above for calendar year initiation estimates.

One important note for both the calendar year and past year estimates of incidence is the relationship between a main substance category and subcategories of substances (e.g., illicit drugs would be a main category and inhalants and marijuana would be examples of subcategories in relation to illicit drugs). Typically, any member of a subcategory is by necessity a member of the main category (e.g., if a respondent is a past month user of a particular drug, then he or she is also a past month user of illicit drugs in general). However, this is not the case with regard to incidence statistics. Because an individual can only be an initiate of a particular substance category (main or sub) a single time, a respondent with lifetime use of multiple substances may not, by necessity, be included as an initiate of a main category, even if he or she were an initiate for a particular subcategory because his or her first initiation of other substances could have occurred earlier.

For more information on calendar year and past year incidence, see Appendix B.4.1 of the 2004 NSDUH final results report (Office of Applied Studies, 2005).

## 9. Suppression of Estimates with Low Precision

Direct survey estimates that were considered to be unreliable due to unacceptably large sampling errors were not reported, but rather were noted by an asterisk (\*). The criterion used for suppressing all direct survey estimates was based on the relative standard error (RSE), which is defined as the ratio of the standard error (SE) over the estimate.

Proportion estimates ( $\hat{p}$ ) within the range  $0 < \hat{p} < 1$ , rates, and corresponding estimated numbers of users were suppressed if

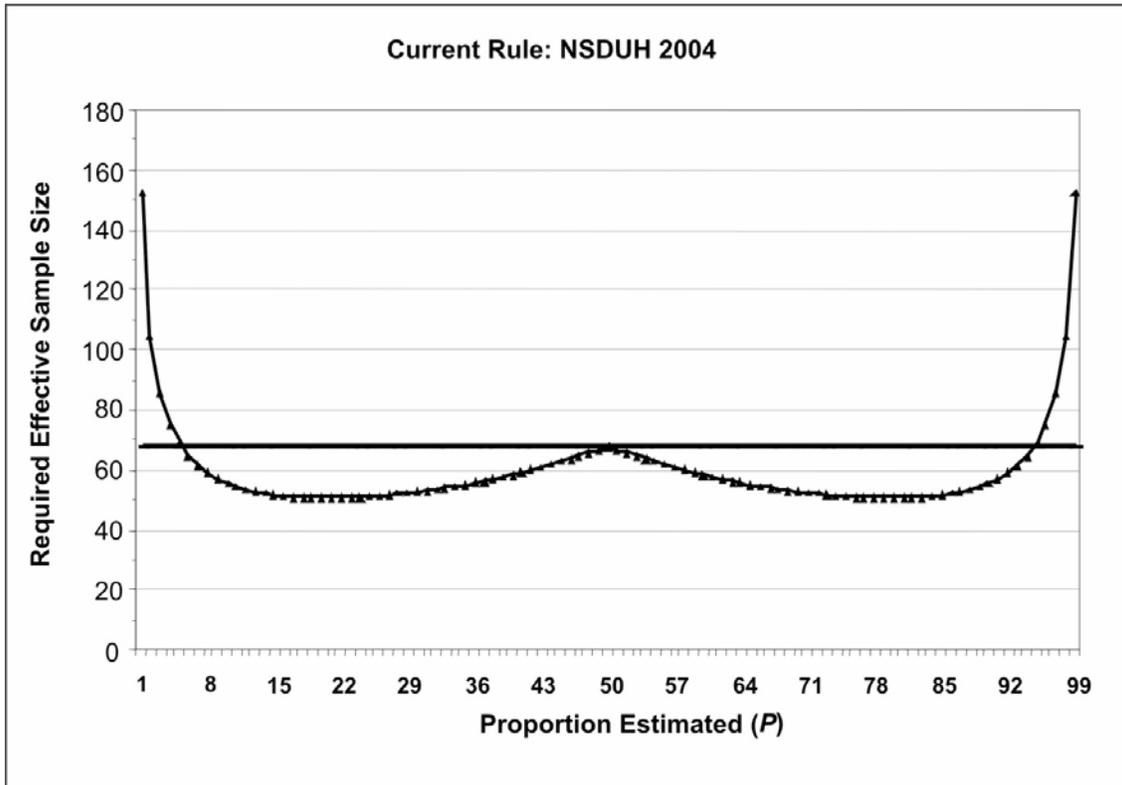
$$\begin{aligned} \text{RSE} [-\ln(\hat{p})] > .175 \text{ when } \hat{p} \leq .5, \\ \text{or} \\ \text{RSE} [-\ln(1 - \hat{p})] > .175 \text{ when } \hat{p} > .5. \end{aligned}$$

Based on a first-order Taylor series approximation of  $\text{RSE} [-\ln(\hat{p})]$  and  $\text{RSE} [-\ln(1 - \hat{p})]$ , the following suppression rule was used for computational purposes:

$$\begin{aligned} \frac{\text{SE}(\hat{p}) / \hat{p}}{-\ln(\hat{p})} > .175 \text{ when } \hat{p} \leq .5, \\ \text{or} \\ \frac{\text{SE}(\hat{p}) / (1 - \hat{p})}{-\ln(1 - \hat{p})} > .175 \text{ when } \hat{p} > .5. \end{aligned}$$

The separate formulas for  $\hat{p} \leq .5$  and  $\hat{p} > .5$  produce a symmetric suppression rule; that is, if  $\hat{p}$  is suppressed,  $1 - \hat{p}$  will be suppressed as well. See Figure 1 for a graphical representation of the required minimum effective sample sizes as a function of the proportion estimated. When  $.05 < \hat{p} < .95$ , the symmetric properties of the rule produce local minimum effective sample sizes at  $\hat{p} = .2$  and again at  $\hat{p} = .8$ , such that an effective sample size of greater than 50 is required; this means that estimates would be suppressed for these values of  $\hat{p}$  unless the effective sample sizes were greater than 50. Within this same interval of  $.05 < \hat{p} < .95$ , a local maximum effective sample size of 68 is required at  $\hat{p} = .5$ . So, to simplify requirements and maintain a conservative suppression rule, estimates of  $\hat{p}$  between .05 and .95 which had effective sample sizes below 68 were suppressed.

**Figure 1. Required Effective Sample as a Function of the Proportion Estimated**



A minimum nominal sample size suppression criterion ( $n = 100$ ) that protects against unreliable estimates caused by small design effects and small nominal sample sizes was employed. Prevalence estimates also were suppressed if they were close to 0 or 100 percent (i.e., if  $\hat{p} < .00005$  or if  $\hat{p} \geq .99995$ ).

Estimates of other totals (e.g., number of initiates), along with means and rates not bounded between 0 and 1 (e.g., mean age at first use and incidence rates) were suppressed if the RSEs of the estimates were larger than .5.

Additionally, estimates of mean age of first use were suppressed if the sample sizes were smaller than 10 respondents; also, the estimated incidence rate and number of initiates were suppressed if they rounded to 0.

The suppression criteria for various NSDUH estimates are summarized in Table 1.

**Table 1. Summary of 2004 NSDUH Suppression Rules**

<b>Estimate</b>	<b>Suppress if:</b>
Prevalence rate, $\hat{p}$ , with nominal sample size, $n$ , and design effect, $deff$	<p>(1) The estimated prevalence rate, <math>\hat{p}</math>, is less than .00005 or greater than or equal to .99995, or</p> <p>(2) <math>\frac{SE(\hat{p})/\hat{p}}{-\ln(\hat{p})} &gt; .175</math> when <math>\hat{p} \leq .5</math>, or</p> $\frac{SE(\hat{p})/(1-\hat{p})}{\ln(1-\hat{p})} > .175$ when $\hat{p} > .5$ , or <p>(3) Effective <math>n &lt; 68</math>, where Effective <math>n = \frac{n}{deff}</math>, or</p> <p>(4) <math>n &lt; 100</math>.</p> <p>Note: The rounding portion of this suppression rule for prevalence rates will produce some estimates that round at one decimal place to 0.0 percent or 100.0 percent but are not suppressed from the tables.</p>
Estimated number (numerator of $\hat{p}$ )	<p>The estimated prevalence rate, <math>\hat{p}</math>, is suppressed.</p> <p>Note: In some instances when <math>\hat{p}</math> is not suppressed, the estimated number may appear as a 0 in the tables; this means that the estimate is greater than 0 but less than 500 (estimated numbers are shown in thousands).</p>
Mean age at first use, $\bar{x}$ , with nominal sample size, $n$	<p>(1) <math>RSE(\bar{x}) &gt; .5</math>, or</p> <p>(2) <math>n &lt; 10</math>.</p>
Incidence rate, $\hat{r}$	<p>(1) The incidence rate, <math>\hat{r}</math>, rounds to less than 0.1 per thousand person years of exposure, or</p> <p>(2) <math>RSE(\hat{r}) &gt; .5</math></p>
Number of initiates, $\hat{t}$	<p>(1) The number of initiates <math>\hat{t}</math>, rounds to fewer than 1,000 initiates, or</p> <p>(2) <math>RSE(\hat{t}) &gt; .5</math></p>



# References

Bowman, K. R., Chromy, J. R., Hunter, S. R., & Martin, P.C.(2005, February). Sample design report. In *2004 National Survey on Drug Use and Health: Methodological resource book* (Section 2, prepared for the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract No. 283-03-9028, Deliverable No. 8, RTI/0208726. Research Triangle Park, NC: RTI International. [Available as a PDF at <http://www.oas.samhsa.gov/nhsda/methods.cfm#2k4>]

Chen, P., Dai, L., Folsom, R., Gordek, H., Grau, E., Shi, W., & Westlake, M. (2005, October). Person-level sampling weight calibration. In *2004 National Survey on Drug Use and Health: Methodological resource book* (Section 3, prepared for the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract No. 283-03-9028, Deliverable No. 26, RTI/0208726). Research Triangle Park, NC: RTI International. [Available as a PDF at <http://www.oas.samhsa.gov/nhsda/methods.cfm#2k4>]

Gfroerer, J., Hughes, A., Chromy, J., Heller, D., & Packer, L. (2004, July). Estimating trends in substance use based on reports of prior use in a cross-sectional survey. In S. B. Cohen & J. M. Lepkowski (Eds.), *Eighth Conference on Health Survey Research Methods: Conference proceedings [Peachtree City, GA]* (DHHS Publication No. PHS 04-1013, pp. 29-34). Hyattsville, MD: National Center for Health Statistics. [Available as a PDF at [http://www.cdc.gov/nchs/data/misc/proceedings\\_hsr2004.pdf](http://www.cdc.gov/nchs/data/misc/proceedings_hsr2004.pdf)]

Greenberg, R. S., Daniels, S. R., Flanders, W. D., Eley, J. W., & Boring, J. R. (1996). *Medical epidemiology*. Norwalk, CT: Appleton & Lange.

Office of Applied Studies. (2002, July). *2001 National Household Survey on Drug Abuse: Incentive experiment: Combined quarter 1 and quarter 2 analysis* (prepared by RTI under Contract No. 283-98-9008). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available as a PDF at <http://www.oas.samhsa.gov/nhsda/methods.cfm#2k1>]

Office of Applied Studies. (2005). *Results from the 2004 National Survey on Drug Use and Health: National findings* (DHHS Publication No. SMA 05-4062, NSDUH Series H-28). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at <http://www.oas.samhsa.gov/p0000016.htm#2k4>]

RTI International. (2004a). *SUDAAN<sup>®</sup> language manual, Release 9.0*. Research Triangle Park, NC: Author.

RTI International. (2004b). *SUDAAN<sup>®</sup> manual v. 9.0 (beta)*. Research Triangle Park, NC: Author.

