The Use of External Data Sources and Ratio Estimation To Improve Estimates of Hardcore Drug Use from the NHSDA

Douglas Wright, Joseph Gfroerer, and Joan Epstein

ABSTRACT

Levels of hardcore drug use have been especially difficult to estimate because of the relative rarity of the behavior, the difficulty of locating hardcore drug users, and the tendency to underreport stigmatized behavior. This chapter presents a new application of ratio estimation, combining sample data from the National Household Survey on Drug Abuse (NHSDA) together with population counts of the number of persons arrested in the past year from the Uniform Crime Report (UCR) and the number of persons in drug treatment programs in the past year from the National Drug and Alcoholism Treatment Unit Survey (NDATUS). The population counts serve as a benchmark accounting for undercoverage and underreporting of hard drug users.

INTRODUCTION

The need for accurate estimates of the size of the so-called hardcore drug-using population is substantial. Studies of hardcore drug users typically include heavy chronic users of drugs such as cocaine and heroin. While there is no standard definition of hardcore drug users, and the authors here employ several alternative measures, this population of heavy drug users is likely to need significant resources for treatment of their drug problems and associated medical and other problems. Hardcore drug users have also been shown to be responsible for a disproportionate amount of crime (Nurco et al. 1991).

This chapter describes a method for estimating the prevalence of hardcore drug use based on the NHSDA in conjunction with outside sources and the methodology of ratio estimation. In ratio estimation, one can often obtain a better estimate of a population total if there is a known population total for a related variable. Then the estimate of the total is $Y. = (y/x)*X$, where y is the variable of interest, x is the

related variable, and X is the known population total for the related variable (Cochran 1977).

Another way of describing this method is to say that it inflates (i.e., gives more weight to) the drug prevalence data from the NHSDA for populations with characteristics that are known to be related to hardcore drug use but are also underestimated. In this case, it is known that NHSDA undercounts arrestees and drug treatment populations, so there is a ratio adjustment of the NHSDA hardcore drug use estimates upward to externally derived counts of arrestees and treatment clients that are believed to be accurate.

In survey sampling theory, ratio estimation is often associated with the desire to improve the precision of an estimate. The ratio estimate will be better, in the sense that it will have a smaller variance, than the simple expansion estimator $Y/ = \acute{\imath}w_i y_i$ that is commonly used when certain conditions are met. Here, Y is the variable of interest, y_i is the value reported by the *i*-th sample case, and w_i is the inverse of the sampling probability for the *i*-th case. (Precision of the estimates is discussed later in this chapter.)

However, in this application, one is less interested in variance reduction and more interested in bias reduction. Ratio estimates have been used for a number of years to adjust for nonresponse and to adjust to known population counts, often based on a census. This application uses known population counts to adjust NHSDA sample estimates for underreporting and undercoverage. Early examples of using ratio estimation to adjust for undercoverage and nonresponse are the Health Interview Survey (Thornberry and Massey 1978) and the 1975 Survey of Scientific and Technical Personnel (Tupek and Richardson 1978), respectively. The NHSDA methodology is described elsewhere (Substance Abuse and Mental Health Services Administration (SAMHSA) 1993*a*, 1994).

BACKGROUND

The major problems with obtaining good estimates of the number of hardcore drug users are locating them and getting them to respond honestly about their usage. Studies have shown that hardcore drug users tend to be more transient, being found in significant numbers outside households (National Institute on Drug Abuse (NIDA) 1994). Other attempts to estimate the number of hardcore drug users

typically have been based on model-based assumptions or nonrandom samples. (See section on comparing with other methods.)

Estimating the number of hardcore drug users has historically been a difficult problem. Household interview surveys such as the NHSDA were not designed for this type of estimation and are believed to be inadequate tools for measuring hardcore drug use because of the low prevalence of the behavior and difficulties in accessing this population. Underreporting (survey participants who do not report their drug use) and undercoverage (inability to roster hardcore drug users) also affect this estimation.

In comparing the results of NHSDA estimates to those from various administrative records systems (e.g., drug treatment program data, parole, probation, or arrest data from the Federal Bureau of Investigation (FBI)), the apparent underreporting of these types of characteristics by the sample respondents has been significant. It should be noted that FBI Index Crimes appear to be fairly well estimated by the NHSDA when allowing for differences due to incarceration (Harrison and Gfroerer 1992).

Research has shown that underreporting of drug use increases as the reference period approaches the present and as the perceived social disapproval increases. This suggests that hardcore drug use is underestimated more than casual drug use. The underestimates could also be the result of undercoverage of the populations with these characteristics (Gfroerer 1993; Turner et al. 1992).

Various methods have been used to estimate hardcore drug prevalence including capture-recapture techniques, truncated Poisson, and modeling methods generally (c.f., Brodsky 1985; Hser et al. 1992; Woodward et al. 1985). These methods have been based primarily on data from administrative records such as treatment admission data, essentially ignoring household survey data. Other methods have combined household survey data with other sources of data (e.g., arrest data) to construct composite estimates of hardcore drug use (Rhodes 1993; Wish 1990-91).

There also has been significant research on various data-collection methods that encourage honest responses to sensitive questions. Such methods include randomized response and nominative techniques (Miller 1985; Warner 1965; Zdep et al. 1979).

Randomized response involves a randomizing device, such as a pair of dice, and two questions, one of which is sensitive ("Have you used heroin in the past year?") and one of which is innocuous (e.g., "Were you born in April?"). The respondent uses the randomizing device to determine which question to answer, and the interviewer records the answer (yes or no), but not the question. This method can be used to estimate the proportion having the sensitive characteristic.

The nominative technique involves first asking the respondent to indicate how many of his or her close friends have the sensitive characteristic (e.g., use of heroin in the past year). The second question is then asked about each of the close friends: "How many of this person's other close friends (besides yourself) know that he (or she) has used heroin?" With these two questions the count of the number of past-year heroin users can be corrected for duplication.

In the following discussion, the focus is on the ratio estimate's ability to reduce bias (in particular, the undercounting of hard drug users in the NHSDA) given a true population value of a related variable. To make the discussion more concrete, the estimation procedure will be applied to four separate, but overlapping, measures of hardcore drug use for 1992: the number of past year users of heroin, weekly users of cocaine in the past year, past-year users who are dependent on some illicit drug, and past-year intravenous drug users.

BASIC METHODOLOGY

The information that the authors wish to use are the count of the number of persons in treatment centers for drug abuse during 1992 from the NDATUS (SAMHSA 1993*b*) and the known count of the number of arrests (for any crime other than minor traffic violations) during 1991 from the FBI Uniform Crime Reporting Program (Maguire et al. 1993).

Using a Single Population Count

Let Nt equal the estimated count of the number who received treatment for a drug problem during 1992 derived from the NDATUS. The count was computed by multiplying the number of treatment slots times the average number of persons treated per year per slot, and includes an adjustment for multiple episodes by the same individual.

Na equals the estimated count of the number of persons arrested during 1991. Na was calculated by taking the latest available FBI Uniform Crime Report estimated number of arrests—14,211,900—and dividing this by the average number of arrests per person arrested calculated from the NHSDA, approximately 1.46, resulting in an estimate of 9,722,671. Based on recent trends, the 1992 estimate would be expected to be slightly higher than the 1991 estimate.

The typical use of a ratio estimate occurs when the outside source fully overlaps the population of interest. For example, if one wanted to estimate the number of students in school and had information on the total number of teachers, then this source overlaps the population of interest since every student has a teacher. Then, if an estimate of the average number of students per teacher were developed, one could multiply the total number of teachers times the average number of students per teacher to obtain an estimate of the total number of students.

The authors' situation is slightly different in that neither the treatment nor arrestee population counts fully overlap the population of hard drug users. However, one can construct counts from the Census Bureau estimates for 1992 that are so precise at the national level that they can be considered population counts. With these data, one can develop counts that cover the population.

Given the Census Bureau estimate of the number of noninstitutionalized persons 12 and older, N = 205,713,000 (for July 1, 1992, the count at the midpoint of data collection for the NHSDA target population), one can form two pairs of counts that cover the population.

Number of persons in treatment: Nt = 1,789,000
Number of persons not in treatment: N-Nt = 203,924,000

Number of persons arrested: Na = 9,722,671
Number of persons not arrested: N-Na = 195,990,329.

From these counts, two estimates of the number of persons using heroin (H in the following equation) during the past year (1992) are possible.

$$H(t) = r(t)*Nt + r(t)^{*}*(N-Nt) = (139,003/834,702) * 1,789,000(171,136/200,656,309) * 203,924,000$$
$$= .166*1,789,000 + .00085*203,924,000$$
$$= 297,922 + 173,922$$
$$= 471,844.$$

$$H(a) = r(a)*Na + r(a)^{*}*(N-Na) = (184,277/4,743,706) * 9,722,671 + (125,861/196,747,306) * 195,990,329$$
$$= .039*9,722,671 + .00064*195,990,329$$
$$= 377,693 + 125,377$$
$$= 503,070$$

where $r(t) = h_t/t$: the estimated rate of hardcore drug use (heroin, in this example) in population Nt. For heroin, it is the estimated number (from the NHSDA sample) in treatment and using heroin in the past year divided by the estimated number (from the sample) in treatment centers for drug use in the past year. (HERREC was the heroin variable used from the 1992 NHSDA file to code if the respondent used heroin in the past year; the treatment variable was the union of three variables TRMTTRMT (received treatment for drug use in a treatment or rehab facility in the past year), TRMTHOSP (received treatment for drug use in a hospital in the past year), and TRMTMHC (received treatment for drug use in a mental health center in the past year.) The treatment variable was defined in this way in order to be as consistent as possible with the data collected in NDATUS.)

The estimated rate of hardcore drug use in population N-Nt is denoted $r(t)^{*}$. For heroin, it is the estimated number (from the sample) not in treatment but using heroin in the past year divided by the estimated number (from the sample) not in treatment centers for drug use during the past year.

The estimated rate of hardcore drug use in population Na is denoted $r(a) = h_a/a$. In this example, it is the estimated number (from the sample) arrested and booked and using heroin in the past year divided by the estimated number (from the sample) arrested and booked in the past year (the union of the variable NOBOOKYR with BKLARCNY, BKBURGL, BKAGASLT, BKSMASLT, BKMVTHFT, BKROB, BKRAPE, BKMURDER, BKARSON BKDRVINF, BKDRUNK, BKDRUG, BKPROS, BKVANDAL, and BKOTHOF).

The estimated rate of hardcore drug use in population N-Na is denoted $r(a)^{*}$. It is the estimated number (from the sample) not arrested and

not booked but using heroin during the past year divided by the estimated number (from the sample) neither arrested nor booked in the past year.

The above estimates can be compared to the standard published estimate of 323,000 for the number of past-year users of heroin in 1992. The standard published estimate is a complex estimator incorporating the inverse of the selection probability, adjustments for household and person-level nonresponse, and a poststratification adjustment to census projections of age by gender by race/ethnicity distributions.

Using Two Known Population Counts

Having the two separate estimates based on treatment and arrests counts raises the question, "Is there an alternate method that would make simultaneous use of both the treatment and arrest counts?"

The ideal situation when one has two variables, such as the number receiving treatment in the past year and the number arrested and booked in the past year, is to use known counts for the interior cells. In other words, one can make consistent estimates that make use of ratio estimation for each of the cells using the following matrix:

	Treatment	No treatment
Arrested	N(11)	N(12)
Not arrested	N(21)	N(22)

where N(11) is the known count of the number in treatment and arrested and booked in the past year and so forth for the other cells. (It is interesting to note that if there is no partial nonresponse for the variables of interest, and estimates are used of the number in each cell based on the NHSDA sample multiplied by the sample cell estimates of prevalence, the NHSDA standard published estimate is obtained. The impetus for using accurate external counts is that the sample counts tend to underestimate counts of those in treatment and those arrested and booked.)

In the earlier description of the calculation for heroin, all that was known at the national level were the marginals Nt, the number in treatment in the past year, and Na, the number arrested and booked in the past year. In the absence of known population counts, exploring

a few alternatives is possible. One alternative is to fill in the interior cell counts by raking the marginals (or iteratively proportionally fitting); for example, by raking based on the least squares solution. However, this solution, often suggested when the initial interior cells exhibit variation due to sampling error, minimizes the change of the interior estimates. This would be inappropriate here because the authors believe the cell estimates are biased downward.

A second alternative, and the one the authors employ, is to use independent sample estimates of one cell to estimate the remaining cells. To estimate one of the interior cells, sample data from the 1990 Drug Services Research Survey (DSRS) was used in conjunction with data from the NHSDA and the estimated marginal count from NDATUS.

DSRS is a national sample survey of treatment centers and records of discharged clients. In 1990, a stratified random sample of 1,183 drug treatment facilities was selected from respondents to the 1990 NDATUS file. In the summer of 1990, a stratified subsample of 118 combined drug and alcohol treatment facilities was selected. In each facility, a sample of client discharge records (discharged during the 12 months between September 1, 1989 and August 31, 1990) was selected. The total sample included over 2,000 records (Bigel Institute for Health Policy 1992).

The aim is to estimate the number of persons in treatment during 1992 who also were arrested and booked in 1992, represented as N(11). Because the number of persons in treatment (Nt) is known, the only task is to estimate the percent of those in treatment who also were arrested and booked in 1992.

From DSRS, one can estimate that the proportion of those in treatment who were ever arrested was 77 percent. This number is based on records having information on arrests and on the number of episodes of treatment in the past year. Approximately 30 percent of the records were missing one or more of these variables. The estimate from the 1990 DSRS survey has been assumed to hold true for the target year, 1992. From the NHSDA, one can estimate that the proportion of those in treatment during the past year and ever arrested who were also arrested and booked in the past year was 52 percent. Therefore, it is necessary to multiply 0.77 by 0.52 times the number in treatment, 1,789,000, to obtain 716,315, the number in treatment and arrested and booked in the past year.

Counts for the interior cells N(11), N(12), N(21), and N(22) that are consistent with the marginal counts used earlier, can now be obtained.

	Treatment	No treatment	Total
Arrested	716,315	9,006,356	9,722,671
Not arrested	1,072,685	194,917,644	195,990,329
Total	1,789,000	203,924,000	205,713,000

The following notation can be used to represent the corresponding ratios (i.e., prevalence rates):

	Treatment	No treatment	Marginal avg.
Arrested	r(11)	r(12)	r(a)
Not arrested	r(21)	r(22)	r(a)*
Marginal avg.	r(t)	r(t)*	r

Then the estimate based on interior cells can be written as:

Estimate = H(t,a) = r(11)*N(11) + r(12)*N(12) + r(21)*N(21) r(22)*N(22).

Returning to the estimation of heroin, the prevalence rates in the cells are:

	Treatment	No treatment	Marginal avg.
Arrested	0.308	0.02100	0.03900
Not arrested	0.092	0.00038	0.00064
Marginal avg.	0.166	0.00085	0.00160

(The grand average is the standard published estimate divided by the population total, 205,713,000.)

The estimate based on all four cells rather than the marginals is 587,966, which was calculated as follows:

H(t,a) = (88,495/287,200)*716,315 + (50,507/547,502)*1,072,685
+ (95,781/4,456,505)*9,006,356
+ (75,354/196,199,804)*194,917,644 = 587,966.

This estimate is larger than either of the estimates based on marginals. By analogy with stratification and poststratification, the interior cells estimate is generally to be preferred to marginal estimates, especially if there are large differences in the usage ratios among the cells.

Two comments about interpretation of the above numbers are in order. First, since the above ratios have been calculated only for those records for which all the variables exist, including the treatment and arrest variables, one would expect the denominators to be somewhat of an underestimate of the independent count totals. Second, notwithstanding that fact, two of the cells are significant underestimates by a factor of 2 or more. For example, the NHSDA estimated number of those arrested and booked in the past year but not in treatment, 4,456,505, is less than half of the independent count of 9,006,356. The third cell, those in treatment but not arrested, is underestimated by a factor slightly less than 2. The last cell, those not in treatment and not arrested in the past year, is very close to the independent count, so that the impact of the ratio estimate on this cell is minimal.

To obtain the most accurate estimate, one should utilize independent population counts that are believed to have the best coverage and are most closely related to the variable of interest. One can also calculate the estimated variance of alternative estimates, preferring the one with the smallest variance, other things being equal.

The mean squared error of the estimate based on the interior cells (given that the interior cells are known counts) will typically be smaller as well, as long as the cell sample sizes are sufficiently large and certain relationships between the numerator and denominator of cell prevalence rates hold (in particular, that the correlation between the numerator and denominator is less than the population coefficient of variation (cv) of the denominator divided by twice the population cv of the numerator; the variances should always be estimated to verify that this is the case).

Assumptions

One major assumption made in the above estimation methodology concerns the accuracy of the estimated ratios given the expectation of underestimation of these hard drug populations from the household sample. A basic assumption being made in these ratios is that both numerator and denominator are being similarly underestimated. This would be the case, for example, if drug users underreport their drug use (or it is undercovered) at the same rate as the treatment population underreports their treatment and arrestees underreport being arrested. Taking the estimate of r(t), the expected value of h_t is assumed to equal $c_{ht}*H_t$, where c_{ht} is a constant and H_t is the true value. Similarly, the expected value of t is assumed to equal c_t*Nt, so that the expected value of r(t) equals approximately H_t/Nt when $c_{ht}=c_t$. (Another possible assumption is that $c_{ht}<c_t$ because some will assert that the least accurate NHSDA coverage is of the heaviest users.)

For the complementary cell in the 2-cell estimate, the numerator of the ratio $r(t)^*$, the number of heroin users who are not in treatment, is probably underestimated. But the sample estimate of the denominator will generally not be an underestimate, so that the impact of ratio estimation on this cell is minimal.

Similarly, in the 4-cell estimate, the cell estimate of those who have not been in treatment or arrested and booked in the past year is very similar to the independent count for that cell. Here also, one expects that the assumption of equal underestimating of the numerator and denominator would not hold. Therefore, the ratio estimate still would underestimate this cell for any measure of hard drug use.

Another issue is the extent to which the authors' sample ratios cover the entire population (noninstitutionalized individuals age 12 and over). By design, the NHSDA only covers those who spend some portion of the year in a household, dormitory, or shelter. It does not cover the population that is composed of persons who are in a residential treatment facility for a full year. These occur relatively infrequently since most treatment is outpatient. For those arrested and booked, the group missed in the household survey is those prisoners who are in prison for the full 12 months because they cannot be in the household sample. The "permanently" homeless (those who do not appear in a household or a shelter for an entire year) also are missing.

To calculate such ratios as r(t), r(a), and others, only those cases that had no missing data for any of the variables mentioned above were used. In other words, each variable was coded 1, 0, or missing. Over three variables there are 27 combinations, but only the 8 combinations of 1s and 0s were used. The implicit assumption is that the item nonresponse is at random. About 1,000 cases had one or more variables missing, but out of a sample of 28,000, this only represents about 3 percent.

It should be noted that the ratio adjustment can be looked at differently, as, for example, $h_t * (N_t/t)$. Looked at this way, the ratio adjustment is a weight adjustment that differs by cell, applied to the usual weight. This simplifies the application of this method and provides the capability of estimating hardcore drug use for population subgroups or for other drugs. Thus, the standard NHSDA estimate Y/ is replaced by $Y/ (adj) = \text{í}w_i(adj)*y_i$, where $w_i(adj)$ is the usual weight adjusted by the appropriate cell ratio.

ESTIMATES OF OTHER MEASURES OF HARDCORE DRUG USE

Below are estimates of other measures of hardcore drug use based on the above methodology. Table 1 displays the prevalence estimates by cell, and table 2 shows the resulting estimates.

Implications of Components of the Estimator

It is instructive to look at the relative size of certain components to gain insight into how the estimate might be improved and where its weaknesses are. Note that in the above assumptions, the levels of the components that make up the ratios r(1) and r(2) are not really important to the size of the estimates H and C, respectively, if one believes that the undercounting of the numerator and denominator are similar. The estimate for H, for example, would be larger if one believes that there has been more underreporting of heroin users in treatment than there has been of the total persons in treatment.

The next components to look at are the denominators of the ratios as compared to the independent counts (actually, partly estimated). The sample-based estimate from the 1992 NHSDA for the number of persons in treatment during the past year is 834,702. The universe estimate from NDATUS (adjusted for certain double-counting and for nonresponse, but

TABLE 1. *Cell estimates for other measures of hardcore drug use (variables in parentheses indicate variable names on the NHSDA public use file).*

Cocaine (COCWKF)	Treatment	No treatment	Marginal avg.
Arrested	0.259	0.0160	0.0310
Not arrested	0.108	0.0020	0.0022
Marginal avg.	0.160	0.0023	0.0031
Dependence on any illicit drug*			
Arrested	0.519	0.0930	0.1180
Not arrested	0.348	0.0066	0.0076
Marginal avg.	0.406	0.0085	0.0102
Past year use of needles (NEDYR3)			
Arrested	0.238	0.0560	0.0670
Not arrested	0.068	0.0014	0.0016
Marginal avg.	0.127	0.0026	0.0032

KEY: * = The number of persons dependent on any illicit drug in the past year is based on an algorithm to approximate the "Diagnostic and Statistical Manual of Mental Disorders," 3d ed. revised (DSM-III-R) criteria (Epstein and Gfroerer 1995). This algorithm combines items on symptoms and problems included in the NHSDA questionnaire to approximate five of the nine DSM-III-R criteria for substance dependence and defines as dependent a person who meets two of the five criteria.

not adjusted for any undercoverage from having a incomplete frame) is 1,789,000.

This would imply that if these two numbers are similarly underestimated, then the NHSDA survey undercounts the NDATUS number by 53 percent (or more). Similarly, the sample-based estimate for the number arrested and booked in the past year is 4,743,706, while the estimate of the number of persons arrested in the past year derived from the UCR is 9,722,671. The sample-based estimate "underestimates" the universe estimate of 9,722,671 by 51 percent. "Underestimates" is put in quotes because the two numbers may not be completely analogous. The sample estimate is of persons arrested and booked, while the universe count is for those arrested. In most arrests, however, if one is arrested, one is usually processed or booked. It is apparent for both of these questions that they are significantly underreported to the NHSDA. The

TABLE 2. *Comparison of ratio estimate to standard estimate for various drugs.*

	Ratio estimates			Standard estimate
Drug usage	Treatment / arrest ratio	Treatment marginal ratio	Arrest marginal ratio	Standard published estimate
Past year heroin	587,966	471,844	503,070	323,000
Weekly use of cocaine	829,017	750,504	742,202	642,221
Past year dependence on any drug	2,869,242	2,467,074	2,635,084	2,104,508
Past year needle use	1,019,165	755,977	960,773	659,292

independent counts of the variables enable one to adjust for this underreporting.

What is the impact of this kind of estimation on other non-hardcore drugs? Generally speaking, it is not as dramatic as with the hardcore drugs. The impact was calculated on use of marijuana in the past year (variable MRJYR). The standard published estimate was 17,400,273, while the ratio estimate based on four cells was 19,461,280. The latter estimate is only 12 percent larger.

The reason for this is that marijuana is used more widely in the population, and most users fall in the no treatment/not arrested cell. Therefore, this cell (13,644,235 users) dominates the estimate. The relative differences in prevalence rates among cells is not as dramatic as with the hardcore drugs. The cell rates are as follows:

	Treatment	No treatment
Arrested	0.91	0.46
Not arrested	0.45	0.07

Precision of Estimates

Estimates of variance were calculated for each of the above estimates using a software package that calculates the variance of complex sample surveys using Taylor series (Research Triangle Institute 1992). The current authors used the ratio estimation procedure with poststratification weights. It was assumed that the independent counts were estimated without error.

The estimated standard error for the ratio estimate of heroin was very similar to the estimated standard error for the standard published estimate, approximately 106,000. Even though the cell count of those who have neither been in treatment nor arrested and booked in the past year is the largest of the four cells, its contribution both to the estimate and to the estimate of variance was relatively small because of the estimated low incidence in that cell. The 95 percent (2 &) confidence interval for the estimate of past-year users of heroin was 587,966 plus or minus 212,000.

For the other variables, the standard errors for the standard NHSDA estimates were similar to those for the ratio estimates. Since the standard errors of the estimates have remained similar while the estimates themselves have increased, the coefficients of variation (the relative precision of the estimates) have been somewhat improved.

COMPARISON WITH ESTIMATES FROM OTHER METHODS

Previous national estimates of hardcore drug use have used widely varying methods. Estimates of heroin prevalence published by NIDA in the 1970s relied on a small number of locally derived prevalence estimates that were projected to the entire nation using available heroin problem indicators available in other locations (Person et al. 1977). These estimates of the number of heroin addicts ranged from 584,000 in 1974 to 420,000 in 1979. However, these are not comparable to estimates of any past-year heroin use because past-year users include both addicts and infrequent users.

A nominative method of estimating heroin prevalence from the NHSDA produced an estimate of 1.9 million past-year heroin users in 1982 (Miller 1985). A recent estimate of 658,000 weekly heroin users in 1990 was derived from a synthetic estimation procedure that involved combining multiple data sources under various assumptions

(Rhodes 1993). This same methodology was used to derive an estimate of 2.1 million weekly cocaine users in 1991. These recent synthetic estimates represent the most rigorous attempts to utilize multiple sources of data in estimating hardcore drug use prevalence.

While there are many differences between the synthetic estimation model and the NHSDA ratio estimation, the large discrepancies in estimates from the two methods are largely explained by the assumptions made regarding the arrestee population. The synthetic model relied heavily on drug prevalence data from the Drug Use Forecasting system (not a representative sample of arrestees), which resulted in an estimate of 1.8 million weekly cocaine users and 500,000 weekly heroin users among arrestees. By contrast, the ratio estimation method relies more heavily on NHSDA drug prevalence data for arrestees, and resulted in an estimated 329,626 weekly cocaine users and 414,265 past-year heroin users among arrestees.

A complete evaluation and comparison of the ratio estimation procedure with other methods of estimating hardcore drug use is beyond the scope of this chapter. However, one can make some overall statements about ratio estimation.

- Ratio estimation does not fully account for underreporting and undercoverage in the NHSDA. In particular, for the population not arrested and not in treatment, the method does not adjust for under- reporting at all. Thus, the authors consider these estimates of hardcore drug use to be improvements on the standard published NHSDA estimates but still conservative. The standard published NHSDA estimates already adjust for under- or overcoverage by age, race/ethnicity, and gender.

- Because ratio estimation can be looked at as an adjustment to the NHSDA analytic weights (which are based on a probability-based sample design), it provides analytic capabilities that are not possible in any of the previously used methods. While other methods essentially focus on obtaining the bottom line estimate of the number of hardcore drug users, by constructing estimates within the framework of the NHSDA data set one can extend the estimation to population subgroups, such as by region, gender, race/ethnicity, and income, taking advantage of the multitude of data collected in the NHSDA. It must be acknowledged, however, that there are limitations to these secondary applications of the ratio estimation procedure that are not yet determined. Because the procedure is designed

to improve national hardcore drug use estimates, it may not be appropriate (without modification) for certain other estimations, such as for some subgroups and for other drug use measures (e.g., casual use).

- The ratio estimation model, as applied in this case, relies primarily on regularly updated and consistently collected data from the NHSDA, NDATUS, and UCR, and a relatively small number of easily understood assumptions. Thus, it is likely to be able to provide more reliable trend information (given constant levels of underreporting) than the previously used methods, which rely more heavily on assumptions that could change over time.

- Because ratio estimation relies primarily on the NHSDA sample design and weighting, it is possible to develop estimates of the variances of ratio-adjusted estimates. This is generally not possible in the methods previously used.

POSSIBLE FUTURE RESEARCH/APPLICATIONS

There are three primary areas for further investigation. One is in the population counts. Another is the assumptions made about the ratios used. The third involves a search for unbiased methods to estimate the ratio.

1. It would be useful to explore the development of more accurate estimates for the four-cell counts or of alternative counts based on different variables. Estimating the counts used in this chapter necessitated using multiple sources to make the counts comparable to what is collected by NHSDA. Generally, this is best accom-plished by coordinating the questions on the NHSDA and other surveys with systems used to develop administrative counts so that the definitions are as consistent as possible. Coordination of item wording among surveys will at a minimum make it possible to compare estimates across surveys. For 1994, the NHSDA question on being in treatment has been changed to agree exactly with the definition used in NDATUS.

Since it is known that age and race are major correlates of the rate of drug usage, another improvement would be to seek a source or a method of estimation that could provide further age/race breakouts to the treatment/arrest cell counts.

2. In the area of assumptions, when possible, one can compare the distributions of persons for a variable used in the cross-classification based on the NHSDA to those of the population frames to see if they are similar. For example, one can compare the distribution of those in treatment from the NHSDA to the distribution of the population values from NDATUS that are available by age, race, and gender.

Another possibility is the introduction of additional weights reflecting the proportion of the year that a person is in treatment or living in a household. This would serve to increase the size of the populations that are not year-round household residents.

3. With respect to the instrument, one could perhaps try to introduce methodology that would result in less undercounting of the variables that form the ratios: heroin, treatment, arrested and booked, and others, possibly using multiplicity methods, nominative techniques, or using some new method such as hair tests (if the methodology proves to be feasible) to confirm drug use or nonuse.

4. Other applications: While population counts were not directly available at the national level for each of the interior cells, similar methodology may be useful for smaller geographic entities such as States, where the interior cell counts may be known. To use this methodology, one would have to conduct a prevalence survey including questions as similar as possible to what is collected in existing population counts of related social indicators. Again, more coordination of item wording between sample surveys and administrative file systems would enhance this kind of estimation.

REFERENCES

Bigel Institute for Health Policy, Brandeis University. *Drug Services Research Survey - Final Report: Phase II.* Final report of activities under National Institute on Drug Abuse contract no. 271-90-8319/1. Waltham, MA: Brandeis University, 1992.

Brodsky, M.D. History of heroin prevalence estimation techniques. In: Rouse, B.; Kozel, N.; and Richards, L., eds. *Self-Report Methods of Estimating Drug Use: Meeting Current*

Challenges to Validity. National Institute on Drug Abuse Research Monograph 57. DHHS Pub. No. (ADM)85-1402. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1985. pp. 94-103.

Cochran, W.G. *Sampling Techniques*. New York: John Wiley & Sons, Inc., 1977.

Epstein, J.F., and Gfroerer, J.C. "Estimating Substance Abuse Treatment Need From a National Household Survey." Paper presented at the 37th International Congress on Drug and Alcohol Dependence, San Diego, 1995.

Gfroerer, J.C. An overview of the National Household Survey on Drug Abuse and related methodological research. *Proceedings of the Survey Research Section of the American Statistical Association*. Washington, DC: American Statistical Association, 1993.

Harrison, L., and Gfroerer, J. The intersection of drug use and criminal behavior—Results from the National Household Survey on Drug Abuse. *Crime Delinquency* 38(4):422-443, 1992.

Hser, Y.I.; Anglin, M.D.; Wickens, T.D.; Brecht, M.L.; and Homer, J. *Techniques for the Estimation of Illicit Drug-Use Prevalence: An Overview of Relevant Issues.* National Institute of Justice Pub. No.(NCJ)133786. Washington, DC: National Institute of Justice, 1992.

Maguire, K.; Pastore, A.L.; and Flanagan, T.J., eds. *Sourcebook of Criminal Justice Statistics*. U.S. Department of Justice, Bureau of Justice Statistics. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1993.

Miller, J.D. The nominative technique: A new method of estimating heroin prevalence. In: Rouse, B.; Kozel, N.; and Richards, L., eds. *Self-Report Methods of Estimating Drug Use: Meeting Current Challenges to Validity*. National Institute on Drug Abuse Research Monograph 57. DHHS Pub. No. (ADM)85-1402. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1985. pp. 104-118.

National Institute on Drug Abuse. *Prevalence of Drug Use in the DC Metropolitan Area Household and Nonhousehold Populations: 1991*. Technical Report 8. Rockville, MD: National Institute on Drug Abuse, 1994.

Nurco, D.N.; Hanlon, T.E.; and Kinlock, T.W. *Recent Research on the Relationship Between Illicit Drug Use and Crimes.* Vol. 9. *Behavioral Sciences and the Law*. New York: John Wiley & Sons, 1991.

Person, P.H.; Retka, R.L.; and Woodward, A.J. *A Method for Estimating Heroin Use Prevalence*. DHEW Publication

No.(ADM)77-439. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1977.

Research Triangle Institute. *SUDAAN Users Manual.* Research Triangle Park, NC: Research Triangle Institute, 1992.

Rhodes, W. Synthetic estimation applied to the prevalence of drug use. *J Drug Issues* 23:297-321, 1993.

Substance Abuse and Mental Health Services Administration. *National Household Survey on Drug Abuse: Population Estimates, 1992*. DHHS Pub. No. (SMA)93-2053. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1993*a*.

Substance Abuse and Mental Health Services Administration. *National Drug and Alcohol Treatment Unit Survey (NDATUS) 1991 Main Findings Report.* DHHS Pub. No. (SMA)93-2007. Rockville, MD: U.S. Department of Health and Human Services, 1993*b.*

Substance Abuse and Mental Health Services Administration. *Preliminary Estimates from the 1993 National Household Survey on Drug Abuse.* Advance report number 7. Rockville, MD: U.S. Department of Health and Human Services, 1994.

Thornberry, O.T, and Massey, J.T. Correcting for undercoverage bias in random digit dialed national health surveys. *Annual Proceedings of the American Statistical Association*. Washington, DC: American Statistical Association, 1978.

Tupek, A.R., and Richardson, W.J. Use of ratio estimates to compensate for nonresponse bias in certain economic surveys. *Annual Proceedings of the American Statistical Association.* Washington, DC: American Statistical Association, 1978.

Turner, C.F.; Lessler, J.T.; and Gfroerer, J.C. *Survey Measurement of Drug Use: Methodological Studies*. DHHS Pub. No. (ADM)92-1929. Rockville, MD: National Institute on Drug Abuse, 1992.

Warner, S.L. Randomized response: A survey technique for eliminating evasive answer bias. *J Am Stat Assoc* 60:63-69, 1965.

Wish, E.D. Drug policy in the 1990s: Insights from new data on arrestees. *Int J Addict* 25:377-409, 1990-1991.

Woodward, J.A.; Bonett, D.G.; and Brecht, M.L. Estimating the size of a heroin abusing population using multiple-recapture census. In: Rouse, B.; Kozel, N.; and Richards, L., eds. *Self-Report Methods of Estimating Drug Use: Meeting Current Challenges to Validity*. National Institute on Drug Abuse Research Monograph 57. DHHS Pub. No. (ADM)85-1402. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1985. pp. 158-171.

Zdep, S.M.; Rhodes, I.N.; Schwartz, R.M.; and Kilkenny, M.J. The validity of the randomized response technique. *Pub Opinion Q* 43:544-549, 1979.

AUTHORS

Douglas Wright, M.S.
Mathematical Statistician

Joseph Gfroerer
Chief, Prevalence Branch

Joan Epstein, M.A.
Health Statistician

Office of Applied Studies
Substance Abuse and Mental Health Services Administration
Room 16C-06
5600 Fishers Lane
Rockville, MD 20857

Click here to go to page 498