Practice of Epidemiology

Alternative Approaches to Assessing Nonresponse Bias in Longitudinal Survey Estimates: An Application to Substance-Use Outcomes Among Young Adults in the United States

Brady Thomas West* and Sean Esteban McCabe

* Correspondence to Dr. Brady Thomas West, Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan Ann Arbor, 4118 Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48104 (e-mail: bwest@umich.edu).

We evaluated alternative approaches to assessing and correcting for nonresponse bias in a longitudinal survey. We considered the changes in substance-use outcomes over a 3-year period among young adults aged 18–24 years ($n = 5,199$) in the United States, analyzing data from the National Epidemiologic Survey on Alcohol and Related Conditions. This survey collected a variety of substance-use information from a nationally representative sample of US adults in 2 waves: 2001–2002 and 2004–2005. We first considered nonresponse rates in the second wave as a function of key substance-use outcomes in wave 1. We then evaluated 5 alternative approaches designed to correct for nonresponse bias under different attrition mechanisms, including weighting adjustments, multiple imputation, selection models, and pattern-mixture models. Nonignorable attrition in a longitudinal survey can lead to bias in estimates of change in certain health behaviors over time, and only selected procedures enable analysts to assess the sensitivity of their inferences to different assumptions about the extent of nonignorability. We compared estimates based on these 5 approaches, and we suggest a road map for assessing the risk of nonresponse bias in longitudinal studies. We conclude with directions for future research in this area given the results of our evaluations.

longitudinal data analysis; multiple imputation; NESARC; National Epidemiologic Survey on Alcohol and Related Conditions; nonignorable nonresponse bias; substance use; survey nonresponse

Analyses of longitudinal data frequently advance scientific understanding of the epidemiology of certain health behaviors. Indeed, a recent search (December 2015) of the *American Journal of Epidemiology* archives revealed 187 publications with the term "longitudinal" in the study title. Unfortunately, all longitudinal studies collecting repeated measures from the same individuals over time are subject to attrition: Some individuals do not provide data in follow-up waves of the study. If the individuals who fail to respond are systematically different from the individuals who do respond in terms of their behaviors and patterns of behaviors over time, then estimates of trends and patterns in the behaviors based only on respondents may be subject to bias. We refer to this bias, or the systematic difference between a respondent-based estimate and a true full-sample value, as nonresponse bias, and this study specifically considered nonresponse bias in longitudinal health surveys.

Assessing the risk of nonresponse bias in longitudinal surveys and possibly adjusting estimates for attrition requires the use of auxiliary variables that are predictive of both the probability of responding at a follow-up wave and the key outcomes being studied over time, including the measures of change that are of research interest [1]. A failure to use auxiliary variables with these 2 important properties can result in misleading estimates of bias risk and adjusted estimates that are biased and inefficient. Longitudinal surveys (relative to

cross-sectional surveys) provide analysts with an advantage in this regard; a wealth of auxiliary information, including measures of key outcome variables at previous waves, can be employed to study attrition patterns and assess the risk of nonresponse bias. However, some methods of estimating nonresponse bias and adjusting estimates may not be optimal for the different mechanisms that ultimately lead to attrition in longitudinal surveys, and this introduces a need for the analyst to assess the sensitivity of their estimates to assumptions about these mechanisms.

We evaluated 5 alternative methods for adjusting longitudinal survey estimates for attrition and assessing the risk of nonresponse bias based on the adjusted estimates. We applied these methods to repeated measures of the most prevalent substance-use behaviors (for alcohol, marijuana, tobacco, and other drug use) and substance-use disorders (as defined by the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* ([2])) collected from a large sample of young adults ($n = 5,199$) in the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). Tobacco, alcohol, marijuana, and other drug use and related substance-use disorders have remained most prevalent among young adults aged 18–24 years over the past several decades in the United States ([3–7]), making unbiased estimation of these trends essential for understanding the effectiveness of programs and policies designed to address this problem. We discuss the advantages and limitations of these alternative approaches, and we provide practicing epidemiologists with a road map for assessing possible nonresponse bias in longitudinal studies.

METHODS

NESARC sample of young adults

In wave 1, the NESARC sample included 5,199 young adults aged 18–24 years. After applying wave-1 sampling weights, these respondents represented a population of young adults that was 50% women, 62% white, 18% Hispanic, 13% African-American, 5% Asian, and 2% Native American (or other). In the second wave of NESARC, 23.9% of the wave-1 respondents did not respond to the follow-up survey request, resulting in a sample of 3,958 young adults measured in both waves. After adjusting the wave-1 sampling weights for this attrition (details below), these 2-wave respondents represented a target population with nearly identical features. More details about the NESARC sample design and data collection methods for both waves 1 and 2 are available elsewhere ([8, 9]).

We assessed patterns of attrition among young adults by performing weighted, design-based cross-tabulation analyses, comparing different groups of young adults based on wave-1 substance-use behaviors and disorders (described below) in terms of nonresponse rates in wave 2. Weighted estimates of nonresponse rates in wave 2 were computed for the different groups, enabling inference about expected response behaviors for the entire US population of young adults in this type of data collection, and the nonresponse rates were compared statistically using design-adjusted Rao-Scott tests of association, taking the stratified multistage cluster sample design of NESARC into account ([9]). These analyses were performed using the -svy: tab- command in Stata, version 14.1 (StataCorp LP, College Station, Texas), and appropriate subpopulation analyses for complex samples were employed to ensure appropriate variance estimation for the estimated rates ([10]).

Measures

We extracted selected demographic and background characteristics from the wave-1 data file, including age, sex, race/ethnicity, educational level (less than high school, high school, some college or higher), and personal income (less than $5,000, $5,000–$14,999, and $15,000 or more, based on estimated tertiles for young adults). We then examined selected substance-use outcomes at each wave (for tobacco, alcohol, marijuana, and other drugs); substance-use disorders; anxiety, mood, or personality disorders; and various measures of substance-abuse treatment utilization. More details regarding these measures are available elsewhere ([11]).

Using data from both waves of NESARC, we developed simple, ordered categorical measures of change in use/disorders for alcohol, tobacco, and marijuana from baseline (wave 1) to follow-up (wave 2). Each measure of change defined the following 4 subgroups, ordered by either the frequency of use or severity of the disorder: 1) no past-year use/disorder at both times; 2) past-year use/disorder at baseline only; 3) past-year use/disorder at follow-up only; and 4) past-year use/disorder at both times.

Alternative adjustment approaches

We evaluated 5 alternative adjustments for attrition that one could perform when analyzing longitudinal survey data, fully accounting for the complex sample design features of NESARC when evaluating each approach ([12]).

Approach 1: no adjustment. Following this approach, we computed weighted estimates of distributions on the categorical change variables in addition to design-adjusted standard errors for the estimated proportions using only cases that responded in both waves and the wave-1 respondent weights (Table [1]). This approach assumes that attrition is ignorable and occurring completely at random, independent of any other observed variables or the measures of change.

Approach 2: adjust wave 1 weights for nonresponse. This approach is similar to approach 1, but it uses the adjusted wave-2 weights provided in the public-use NESARC data (Table [1]). In general, survey organizations adjust initial respondent weights for attrition by first predicting the probability of response at the follow-up wave as a function of information collected in the baseline wave and then multiplying the initial respondent weights by the inverse of the predicted response probability (or some function of it) for a given individual ([12–14]). While weighting adjustments are typically not performed by secondary analysts, interested readers can see Valliant et al. ([12]) for practical advice on this process.

This weight adjustment approach assumes that attrition is ignorable and occurring at random, conditional on all of the

Table 1. Factors Accounted for by the Final Respondent Weights in the National Epidemiologic Survey on Alcohol and Related Conditions Surveys in Wave 1 (2001–2002) and Wave 2 (2004–2005), United States

Weighting Component	Respondent Weight	
	Wave 1	**Wave 2**
Sampling	Unequal probability of selection into the wave-1 sample	Not applicable
Nonresponse	Wave-1 response rates in subgroups defined by metropolitan statistical area status, race/ethnicity, age, and region	Wave-2 response rates in subgroups defined by combinations of geographic region, age, and lifetime mood, anxiety, or personality disorders (from wave 1)
Calibration	2000 US Census distributions by region, age, sex, race, and ethnicity	2000 US Census distributions by region, age, sex, race, and ethnicity

covariates used to perform the adjustments (Table 1) but independent of actual measures of change (i.e., within a subgroup formed by these covariates, those who drop out of the study represent a random sample of all cases in that subgroup).

Approach 3: MI assuming ignorable attrition. In longitudinal studies, multiple imputation (MI) procedures (15) provide an attractive alternative to standard weighting procedures, as these procedures enable predictions of missing values on individual variables that draw on the predictive power of the most relevant correlates of a given survey item with missing data. Imputation can be an especially effective procedure for correcting nonresponse bias in panel surveys when variables of interest are correlated across waves, and outcomes at previous waves can be used to impute values at later waves (16). In longitudinal surveys, the presence of several auxiliary variables that help to improve the predictive power of imputation models can increase the efficiency of MI methods and related estimates of rates and trends. NESARC provides analysts with public-use data sets that contain a wealth of relevant auxiliary information from both waves for many substance-use outcomes of interest.

We used the sequential regression imputation technique (15), implemented in the -mi impute chained- procedure in Stata, to impute missing values on the substance-use and disorder variables in wave 2. See Web Appendix 1 (available at http://aje.oxfordjournals.org/) for the annotated code. Specifically, we used chained multinomial logistic regression equations in -mi impute- (10 imputations, 5 burn-in iterations each) and the full NESARC data set (including older adults, to maintain the design features of the full NESARC sample). We could not simultaneously impute the use and disorder variables for a given substance because they are perfect functions of each other; we therefore performed 2 sets of MI analyses. The imputation models included as fully observed covariates age, sex, race/ethnicity, income, and education; wave-1 measures of past-year use of tobacco, alcohol, and marijuana (for the wave-2 past-year use variables only); wave-1 measures of past-year nicotine dependence, alcohol-use disorder, and marijuana-use disorder (for the wave-2 disorder variables only); and wave-1 indicators of any anxiety, mood, or personality disorders.

In each of the 10 imputed data files, we used the imputed values of the substance-use and disorder variables from wave 2 to assign each case to one of the 4 categories on each of the 6 change variables. We then used design-based MI estimation to compute weighted MI estimates of the proportions (and their MI standard errors) strictly for the subpopulation of young adults aged 18–24 years. Like approach 2, this approach assumes that attrition is ignorable and occurring at random, conditional on the covariates used in the imputation models.

Prior research has suggested that the fraction of missing information (FMI), an important diagnostic statistic arising from MI analyses, may be able to provide analysts with a sense of when a missing data mechanism is nonignorable (17). In these cases, the probability of having missing data on a given item depends on the item itself, even after conditioning on other observed information. Nonignorable missing data mechanisms therefore prevent imputations based on other observed variables (including baseline measures of key survey outcomes in panel surveys) from fully correcting for nonresponse bias. The FMI indicates what proportion of the total variance in an MI estimate is due to between-imputation variance or to uncertainty in the parameter of interest across imputed data sets. More uncertainty would suggest that predictions of missing values for the variables used to compute an estimate of interest were highly variable across the imputations, meaning that the imputation models used to predict missing values were poor, and the auxiliary variables did not have the ability to predict the variables with missing data. Less uncertainty would suggest the opposite: that predictions were stable across the imputed data sets, with reduced uncertainty.

While the FMI is often lower than the simple nonresponse rate for a given variable, given that predictions based on an MI process have the ability to recover missing information in the variables of interest, an inability to recover this information using the auxiliary variables is still possible. A recent simulation study suggested that when the FMI is larger than the nonresponse rate, a missing data mechanism may be nonignorable (18). We consider this possibility in the context of young adults in NESARC.

Approach 4: adjusted estimates based on selection model predictions. Selection models enable one to compute unbiased estimates of the parameters in a substantive model of interest from a "selected sample" that may arise from some larger overall sample according to a nonrandom selection mechanism. This selection mechanism could potentially introduce bias in estimates of the model parameters depending on the features of the selected sample, and selection models provide a means of correcting this bias. In the case of longitudinal surveys, individuals responding at all

time points may represent a nonrandom subsample of the original overall sample, where individuals who drop out of the survey may have unique values on the variable of interest, even when conditioning on other covariates (i.e., nonignorable selection).

Selection models involve 2 dependent variables: the substantive variable of interest and the response (or selection) indicator. Both variables are modeled simultaneously using covariates available for the full sample (which are more readily available in longitudinal surveys). Ideally, the model should include several covariates strongly related to both outcomes, and "instrumental variables" related only to the response outcome; simulation studies have shown that adjustments based on the selection model under these settings do the best job of correcting nonresponse bias (19). The sample of respondents in a longitudinal survey may suffer from "selection on unobservables," where the normally distributed random error that determines whether a case responds (after conditioning on available covariates) is correlated with the normally distributed random error that ultimately determines values on a variable of interest (again after conditioning on available covariates); see McGovern et al. (20) for discussion regarding this assumption of bivariate normality. The underlying error terms for these 2 dependent variables may be correlated (i.e., the substantive variable is correlated with the probability of responding), and the stronger this correlation, the stronger the selection bias (and the stronger the correction in the estimated parameters of the substantive model). For more on selection models, please see Van de Ven and Van Pragg (21) or De Luca and Perotti (22).

For the present application, we followed the approach used by Bärnighausen et al. (23) and Clark and Houle (19, 24). We first estimated an ordered probit selection model for each of our 4-category change outcomes of interest (implemented in the design-based analysis framework in Stata with -svy: heckoprobit-), using the same covariates considered in the imputation models described above in the substantive models, and instrumental variables derived from the wave-2 response models reported by McCabe and West (25) for NESARC. (See the Stata syntax in Web Appendix 1 for detailed discussion; the sociodemographics were generally important predictors in both models.) We then computed predicted probabilities for each of the 4 ordered change categories based on the fitted model, conditional on not responding in wave 2 (facilitated by the -predict, pcond0- postestimation command in Stata). Given these predicted probabilities for nonrespondents in wave 2 based on the selection model and the actual change outcomes for wave-2 respondents, we then computed adjusted, weighted estimates of the predicted probability of being in each change category for the full sample, following Section 3.4 of Clark and Houle (19), in addition to linearized standard errors for the estimates.

Unlike approaches 1, 2, and 3, this selection model approach allows for attrition to be nonignorable. We consider estimated correlations of the random error terms in the selection and outcome equations when interpreting our results to assess the extent of the nonignorability.

Approach 5: MI assuming nonignorable attrition. For the fifth approach, we used PROC MI in SAS, version 9.4 (SAS

Institute, Inc., Cary, North Carolina), to implement the same sequential regression procedure used for approach 3 but employing pattern-mixture models (26–28) to accommodate potential nonignorable missing data mechanisms for the wave-2 variables. The newest version of PROC MI includes a missing-not-at-random option that implements pattern-mixture model approaches for both continuous and categorical variables and enables analysts to adjust imputed values in a way that reflects possible differences between respondents and nonrespondents, allowing for sensitivity analyses (29, 30). The SAS code in Web Appendix 1 indicates the adjustments that were considered. Like approach 4, this approach assumes nonignorable attrition.

Previous simulation studies. Numerous simulation studies have shown that adjusted estimates assuming ignorable missing data mechanisms can be substantially biased if the underlying mechanism is in fact nonignorable. The effectiveness of pattern-mixture model approaches has been demonstrated for both continuous and categorical variables in several cases (26–28), and additional work (19, 31) has evaluated the ability of adjustments based on carefully specified selection models to eliminate bias due to nonignorable nonresponse. Furthermore, a recent simulation study indicated that MI methods tend to outperform weighting adjustments for nonresponse when the imputation models are well-specified (32). In the Discussion section, we expand on the importance of examining the sensitivity of longitudinal survey estimates to assumptions about the attrition mechanism and the approach used.

RESULTS

Predictors of nonresponse in wave 2

Table 2 presents estimates of weighted nonresponse rates in wave 2 as a function of baseline substance-use behaviors and disorders. The design-based Rao-Scott tests of association suggest that alcohol use in wave 1 is a strong predictor of nonresponse in wave 2, with an estimated 26% of lifetime abstainers not responding to the follow-up survey request and significantly fewer individuals in the subgroups using alcohol in wave 1 not responding in wave 2 ($P < 0.05$). Table 2 also shows that none of the disorders in wave 1 appeared to be strong predictors of nonresponse in wave 2 in these analyses, with a possible exception being those diagnosed with a marijuana-use disorder. Finally, the results in Table 3 are largely consistent with the results in Table 2, suggesting that alcohol abstainers tend to have larger rates of nonresponse than those with no lifetime treatment for alcohol use and those with some treatment ($P < 0.05$). These initial descriptive results suggest that estimates of change in behaviors for the young adults may be biased in the direction of alcohol users.

Evaluation of the 5 adjustment approaches

In Table 4, we find that the weighting-adjustment method does not yield substantially different estimates or standard errors relative to the use of wave-1 weights without any

Table 2.   Estimated Nonresponse Rates in Young Adults (Aged 18–24 Years) at 3-Year Follow-up as a Function of Baseline Substance-Use Behaviors and Disorders, National Epidemiologic Survey on Alcohol and Related Conditions, United States, 2001–2002 and 2004–2005

Baseline Substance-Use Outcomes	No. of Respondents	Weighted Nonresponse Rate, %[a]	Rao-Scott Test	
			Test Statistic[a]	P Value
Tobacco use				
No lifetime tobacco use	3,402	22.34	$F(1.96, 127.71) = 0.012$	0.988
Use prior to past year only	182	22.46		
Use in the past year	1,615	22.55		
Alcohol use				
No lifetime alcohol use	1,234	25.69	$F(1.99, 129.59) = 3.608$	0.030
Alcohol use prior to past year only	414	19.88		
Alcohol use in the past year	3,551	21.69		
Marijuana use				
No lifetime marijuana use	3,791	22.26	$F(1.96, 127.51) = 0.238$	0.784
Marijuana use prior to past year only	723	22.15		
Marijuana use in the past year	635	20.76		
Any illicit drug use including marijuana				
No lifetime drug use	3,719	22.97	$F(1.98, 128.66) = 0.692$	0.501
Other drug use prior to past year only	694	21.87		
Other drug use in the past year	781	20.67		
Nicotine dependence				
No lifetime nicotine dependence	4,279	23.00	$F(1.97, 128.13) = 1.739$	0.180
Nicotine dependence prior to past year only	103	17.02		
Nicotine dependence in the past year	817	20.32		
Alcohol-use disorder				
No lifetime alcohol-use disorder	3,853	22.67	$F(1.98, 128.45) = 1.418$	0.246
Alcohol-use disorder prior to past year only	481	18.82		
Alcohol-use disorder in the past year	865	23.33		
Marijuana-use disorder				
No lifetime marijuana-use disorder	4,545	22.96	$F(1.94, 125.97) = 2.674$	0.075
Marijuana-use disorder prior to past year only	366	17.12		
Marijuana-use disorder in the past year	288	21.44		

[a] All weighted estimates and Rao-Scott tests use design information from wave 1 (2001–2002).

adjustments for attrition. We also note that while the MI estimates assuming ignorable attrition are generally quite similar to the weighted estimates, they tend to have higher efficiency (i.e., lower standard errors) than the estimates based on the adjusted wave-2 weights. This is expected, because these imputations are drawing on the strongest predictors of each individual variable, unlike the weights. When comparing the estimated FMI values in Table 4 with the overall wave-2 response rates for each of the individual measures of change, we see that one of the 6 categorical change variables had an estimated proportion where the FMI value was greater than the unweighted response rate for that particular variable (a single proportion for alcohol-use

disorder), and other FMI values were approaching the wave-2 response rates. These findings suggest that some of the missing-data mechanisms may in fact be nonignorable (18), motivating the examination of the approaches allowing for nonignorable attrition.

When evaluating the adjusted estimates based on the selection-model approach, we start to see some evidence of nonignorable attrition possibly affecting the estimates. For past-year tobacco use, the estimate of the correlation parameter in the selection model (which we refer to as ρ) was minimal (−0.047), suggesting minimal selection bias based on unobservable factors. Accordingly, the adjusted estimates were similar to those found using the adjustments assuming

Table 3.   Estimated Nonresponse Rates in Young Adults (Aged 18–24 Years) at 3-Year Follow-up as a Function of Baseline Drug-Treatment Utilization, National Epidemiologic Survey on Alcohol and Related Conditions, United States, 2001–2002 and 2004–2005

Baseline Drug-Treatment Utilization	No. of Respondents	Weighted Nonresponse Rate, %[a]	Rao-Scott Test	
			Test Statistic[a]	P Value
Any alcohol treatment				
No lifetime alcohol treatment (abstainer)	1,234	25.69	$F(1.99, 129.35) = 3.476$	0.034
No lifetime alcohol treatment (drinker)	3,769	21.30		
History of alcohol treatment	162	23.90		
AA or other 12-step meeting				
No lifetime AA/12-step meeting (abstainer)	1,234	25.69	$F(1.93, 113.80) = 0.496$	0.604
No lifetime AA/12-step meeting (drinker)	62	19.58		
History of AA/12-step meeting	99	26.88		
Any drug treatment				
No lifetime drug treatment (abstainer)	3,719	22.97	$F(1.96, 127.43) = 0.785$	0.456
No lifetime drug treatment (drug user)	1,343	21.01		
History of drug treatment	131	22.73		
NA or other 12-step meeting				
No lifetime NA/12-step meeting (abstainer)	3,719	22.97	$F(1.99, 129.07) = 0.664$	0.516
No lifetime NA/12-step meeting (drug user)	66	27.38		
History of NA/12-step meeting	65	17.61		

Abbreviations: AA, Alcoholics Anonymous; NA, Narcotics Anonymous.

[a] All weighted estimates and Rao-Scott tests use design information from wave 1 (2001–2002).

ignorable attrition. For past-year alcohol use, we found a slightly larger positive value of ρ (0.101), suggesting that individuals in higher-ordered categories were more likely to respond after accounting for all of the covariates in the substantive and response models. This matches our descriptive analysis, but we note that the selection-model estimates provide a better sense of the possible bias introduced, given that there was still a correlation between this change outcome and response after taking the available covariates used in the adjustments assuming ignorable attrition into account. For past-year marijuana use, we found a negative and more substantial estimate of ρ (−0.411), suggesting greater nonresponse bias in the opposite direction: Individuals in higher-ordered categories were less likely to respond. This was borne out in the adjusted estimates based on the selection model.

For the measures capturing change in dependence, the estimates of ρ were always negative and varied from −0.278 (marijuana dependence) to −0.531 (alcohol dependence), suggesting that individuals in higher-ordered categories were once again less likely to respond after accounting for the covariates in the selection model. This nonignorable selection resulted in evidence of fairly substantial bias in the change estimates, and this was especially true for alcohol dependence, where adjustments assuming ignorable attrition would overstate the number of young adults not dependent on alcohol at both points. This finding is also consistent with

the relatively high FMI values for the alcohol-dependence proportions. Similar patterns were seen for the other 2 change outcomes regarding dependence, but the nonresponse bias did not appear to be as substantial (given the lower estimated ρ values).

In general, we also found that the selection-model estimates are more efficient than the adjusted estimates assuming ignorable attrition, but this could be due to the fact that uncertainty in the predicted probabilities for the nonresponding cases was not fully accounted for in these standard errors. We discuss this point in more detail in the Discussion section.

Finally, we note that the estimates based on the pattern mixture–model imputation approach were similar to the estimates based on the selection models and generally even more efficient. Alternative choices of adjustments to the fitted generalized logit models, which were used to compute predictions for the nonrespondents allowing for different levels of nonignorability (29, 30), were not found to substantially affect the adjusted estimates. The SAS code in Web Appendix 1 indicates how to consider alternative adjustments.

DISCUSSION

The results of this study suggest that selected longitudinal estimates of change in common substance-use behaviors

Table 4. Estimated Distributions of Changes in Substance-Use Behaviors and Disorders From Baseline (2001–2002) to Follow-up (2004–2005) for Young US Adults (Aged 18–24 Years), Using Complete Cases That Responded in Both Waves of the National Epidemiologic Survey on Alcohol and Related Conditions or After Alternative Approaches of Imputing Missing Wave-2 Values

Changes in Substance-Use Behaviors and Disorders	Weighted Complete-Case Estimate				Multiple Imputation Estimate Using Wave-1 Weights		Wave-2 Nonresponse Rate, %/FMI from MI, %	Selection Model Estimate Using Wave-1 Weights[a]		Multiple Imputation Estimate Using Pattern-Mixture Model Approach and Wave-1 Weights	
	Wave-1 Weights		Wave-2 Weights (Wave-1 Weights Adjusted for Attrition in Wave 2)								
	Estimate, %	LSE	Estimate, %	LSE	MI Estimate, %	MI-LSE[b]		Estimate, %	LSE	MI Estimate, %	MI-LSE
Past-year tobacco use (*n* = 3,958 complete cases)											
No past-year tobacco use, both times	59.61	1.11	60.32	1.12	59.87	1.07	23.87/4.26	60.01	1.02	60.73	0.80
Past-year tobacco use at baseline only	7.69	0.52	7.59	0.51	7.34	0.49	23.87/16.21	7.27	0.41	7.06	0.24
Past-year tobacco use at follow-up only	6.96	0.56	6.93	0.54	6.63	0.51	23.87/18.72	5.98	0.43	5.78	0.69
Past-year tobacco use at both times	25.74	1.04	25.16	1.04	26.16	0.94	23.87/4.43	26.74	0.91	26.43	0.46
Past-year alcohol use (*n* = 3,958 complete cases)											
No past-year alcohol use, both times	14.51	0.92	14.92	0.91	15.71	0.85	23.87/4.31	16.68	0.84	16.54	0.51
Past-year alcohol use at baseline only	6.87	0.43	7.00	0.44	6.73	0.43	23.87/20.80	6.92	0.33	8.58	1.09
Past-year alcohol use at follow-up only	14.04	0.68	14.25	0.69	13.50	0.59	23.87/9.00	12.44	0.51	12.67	0.52
Past-year alcohol use at both times	64.58	1.27	63.83	1.27	64.06	1.13	23.87/2.98	63.96	1.09	62.21	1.15
Past-year marijuana use (*n* = 3,926 complete cases)											
No past-year marijuana use, both times	78.98	0.88	79.34	0.86	79.70	0.83	24.49/7.96	76.79	0.74	79.03	0.42
Past-year marijuana use at baseline only	7.05	0.55	6.85	0.53	6.57	0.48	24.49/8.50	7.09	0.43	6.49	0.33
Past-year marijuana use at follow-up only	7.38	0.55	7.43	0.54	6.93	0.50	24.49/21.33	7.94	0.42	7.59	0.35
Past-year marijuana use at both times	6.59	0.52	6.38	0.50	6.80	0.48	24.49/8.57	8.18	0.47	6.89	0.31
Past-year nicotine dependence (*n* = 3,958 complete cases)											
No past-year nicotine dependence, both times	73.10	0.94	73.85	0.94	73.81	0.94	23.87/3.95	70.68	0.80	72.41	0.49
Past-year nicotine dependence at baseline only	7.51	0.52	7.28	0.50	6.81	0.48	23.87/9.56	7.71	0.40	6.41	0.25
Past-year nicotine dependence at follow-up only	9.25	0.62	9.18	0.62	9.00	0.57	23.87/10.73	9.58	0.49	10.41	0.40
Past-year nicotine dependence at both times	10.14	0.62	9.69	0.61	10.38	0.59	23.87/6.24	12.03	0.54	10.77	0.31
Past-year alcohol-use disorder (*n* = 3,958 complete cases)											
No past-year alcohol-use disorder, both times	69.94	1.09	70.45	1.06	70.18	1.03	23.87/12.18	65.10	0.85	67.25	0.53
Past-year alcohol-use disorder at baseline only	9.34	0.61	9.11	0.58	8.90	0.55	23.87/15.39	9.60	0.47	7.99	0.31
Past-year alcohol-use disorder at follow-up only	11.90	0.68	11.85	0.67	11.45	0.69	23.87/27.35[c]	13.02	0.53	14.38	0.44
Past-year alcohol-use disorder at both times	8.82	0.56	8.59	0.55	9.47	0.58	23.87/14.11	12.28	0.52	10.38	0.32

Table continues

Table 4. Continued

Changes in Substance-Use Behaviors and Disorders	Weighted Complete-Case Estimate				Multiple Imputation Estimate Using Wave-1 Weights		Wave-2 Nonresponse Rate, %/FMI from MI, %	Selection Model Estimate Using Wave-1 Weights[a]		Multiple Imputation Estimate Using Pattern-Mixture Model Approach and Wave-1 Weights	
	Wave-1 Weights		Wave-2 Weights (Wave-1 Weights Adjusted for Attrition in Wave 2)								
	Estimate, %	LSE	Estimate, %	LSE	MI Estimate, %	MI-LSE[b]		Estimate, %	LSE	MI Estimate, %	MI-LSE
Past-year marijuana-use disorder (n = 3,958 complete cases)											
No past-year marijuana-use disorder, both times	90.06	0.62	90.28	0.60	90.25	0.57	23.87/6.24	89.29	0.52	89.30	0.41
Past-year marijuana-use disorder at baseline only	4.21	0.38	4.09	0.36	3.92	0.34	23.87/4.82	4.16	0.30	3.66	0.18
Past-year marijuana-use disorder at follow-up only	3.80	0.39	3.81	0.39	3.69	0.36	23.87/15.89	4.03	0.30	4.64	0.36
Past-year marijuana use-disorder at both times	1.93	0.28	1.82	0.26	2.14	0.27	23.87/7.85	2.52	0.25	2.40	0.16

Abbreviations: FMI, fraction of missing information; LSE, linearized standard error; MI, multiple imputation.

[a] Following the approach of Bärnighausen et al. (23) and Clark and Houle (19, 24), with n = 5,199.

[b] MI-LSE: the standard error of the multiple imputation estimate, where standard errors for the weighted estimates computed using each imputed data set were estimated using linearization to account for the complex sampling features.

[c] The FMI value was greater than the unweighted wave-2 response rate, suggesting nonignorable attrition (18).

and disorders for young adults from NESARC may be subject to nonignorable nonresponse bias. Based on these results, we recommend that epidemiologists analyzing longitudinal survey data adopt the following approach for sensitivity analysis:

1. Generate the longitudinal survey estimates using the weights provided in a given data set (if available), which are generally adjusted for varying attrition rates in different baseline subgroups (assuming ignorable attrition).

2. Impute missing values at the follow-up wave for nonrespondents multiple (e.g., 10) times, using the best possible correlates of the outcomes of interest from the baseline wave (and again assuming ignorable attrition), and then compute (weighted) MI estimates of distributions on the change measures of interest. In addition, compute the FMI for each of the longitudinal survey estimates based on the multiply imputed data sets, and compare these FMI estimates with the unweighted attrition rates in the follow-up wave. If the FMI estimates are larger than the attrition rates, then the estimates may be subject to nonignorable nonresponse bias (18).

3. Impute missing values at the follow-up wave for nonrespondents using carefully specified selection models (19). Compare weighted estimates based on these imputations to earlier estimates assuming ignorable missing data mechanisms. Report the estimated correlation of the residuals in the selection model with the residuals in the substantive model as evidence of the extent of nonignorable nonresponse bias.

4. Repeat step 3, using pattern-mixture models (which also allow one to specify a range of possible assumptions about the extent of nonignorability), and examine the sensitivity of estimates and inferences to these assumptions.

5. Report a range of possible estimates of the distribution on a change variable of interest under different assumptions about the attrition mechanism. Provide confidence intervals for each of the parameters being estimated under the different approaches.

We believe that this sensitivity-analysis approach will yield a better sense of the possible nonresponse bias in longitudinal survey estimates, which are often based on adjustment methods that assume ignorable attrition. We provide clearly annotated code implementing each of the recommended approaches above in Web Appendix 1.

Given the results in this study, we feel that future research needs to develop methods for incorporating selection models into sequential regression imputation procedures. Tools making this possible will make it easier for data analysts to assess the possibility of nonignorable nonresponse bias while also accounting for the uncertainty in predictions based on a fitted selection model (which was not done in this study). Furthermore, additional evaluation is needed of "doubly robust" methods of modeling survey response and an outcome of interest simultaneously when one or both models may be misspecified (32–35).

More generally, we recommend that research presenting analyses of longitudinal survey data consider the sensitivity analysis approach outlined above. Replications of this

approach in various contexts and for different subject matter will help to ensure that this type of approach for assessing possible nonresponse bias becomes a more standard analytical tool for epidemiologists. More replications of this approach will also prevent the publication of potentially incorrect inferences if nonignorable nonresponse bias is truly a problem in certain longitudinal surveys.

REFERENCES

1. Little RJA, Vartivarian S. Does weighting for nonresponse increase the variance of survey means? *Surv Methodol.* 2005; 31(2):161–168.
2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders.* 4th ed. Washington, DC: American Psychiatric Association; 1994.
3. Compton WM, Grant BF, Colliver JD, et al. Prevalence of marijuana use disorders in the United States: 1991–1992 and 2001–2002. *JAMA.* 2004;291(17):2114–2121.
4. Grant BF, Dawson DA, Stinson FS, et al. The 12-month prevalence and trends in DSM-IV alcohol abuse and dependence: United States, 1991–1992 and 2001–2002. *Drug Alcohol Depend.* 2004;74(3):223–234.
5. Hasin DS, Saha TD, Kerridge BT, et al. Prevalence of marijuana use disorders in the United States between 2001–2002 and 2012-2013. *JAMA Psychiatry.* 2015;72(12): 1235–1242.
6. Johnston LD, O'Malley PM, Bachman JG, et al. *Monitoring the Future National Survey Results on Drug Use, 1975–2014. Volume II: College Students and Adults Ages 19–55.* Ann Arbor, MI: Institute for Social Research, University of Michigan; 2015.
7. Substance Abuse and Mental Health Services Administration. *Results From the 2013 National Survey on Drug Use and Health: Summary of National Findings.* Rockville, MD: Substance Abuse and Mental Health Services Administration; 2014. (National Survey on Drug Use and Health Series H-48) (DHHS publication no. (SMA) 14-4863).
8. Grant BF, Kaplan KD. *Source and Accuracy Statement for the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC).* Rockville, MD: National Institute on Alcohol Abuse and Alcoholism; 2005.
9. Grant BF, Kaplan KD, Shepard K, et al. *Source and Accuracy Statement for Wave 1 of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC).* Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism; 2003.
10. West BT, Berglund P, Heeringa SG. A closer examination of subpopulation analysis of complex sample survey data. *Stata J.* 2008;8(4):520–531.
11. National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health. *Alcohol Use and Alcohol Use Disorders in the United States, a 3-Year Follow-up: Main Findings From the 2004–2005 Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC).* US Alcohol Epidemiologic Data Reference Manual, Volume 8, Number 2. Bethesda, MD: National Institutes of Health; 2010. (NIH Publication No. 10-7677).
12. Valliant R, Dever JA, Kreuter F. *Practical Tools for Designing and Weighting Survey Samples.* New York, NY: Springer; 2013.
13. Oh HL, Scheuren FS. Weighting adjustments for unit nonresponse. In: Madow WG, Olkin I, Rubin DB, eds. *Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies.* New York, NY: Academic Press; 1983: 143–184.
14. Ekholm A, Laaksonen S. Weighting via response modeling in the Finnish Household Budget Survey. *J Off Stat.* 1991;7(3): 325–337.
15. Raghunathan TE, Lepkowski JM, Van Hoewyk J, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol.* 2001; 27(1):85–95.
16. Kalton G. Handling wave nonresponse in panel surveys. *J Off Stat.* 1986;2(3):303–314.
17. Wagner J. The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opin Q.* 2010; 74(2):223–243.
18. Nishimura R, Wagner J, Elliott MR. Alternative indicators for the risk of non-response bias: a simulation study. *Int Stat Rev.* 2016;84(1):43–62.
19. Clark SJ, Houle B. *Evaluation of Heckman selection model method for correcting estimates of HIV prevalence from sample surveys (via realistic simulation)* [working paper no. 120]. Seattle, WA: Center for Statistics and the Social Sciences, University of Washington; 2012.
20. McGovern ME, Bärnighausen T, Marra G, et al. On the assumption of bivariate normality in selection models: a copula approach applied to estimating HIV prevalence. *Epidemiology.* 2015;26(2):229–237.
21. Van de Ven WPMM, Van Pragg BMS. The demand for deductibles in private health insurance: a probit model with sample selection. *J Econom.* 1981;17(2):229–252.
22. De Luca G, Perotti V. Estimation of ordered response models with sample selection. *Stata J.* 2011;11(2):213–239.
23. Bärnighausen T, Bor J, Wandira-Kazibwe S, et al. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology.* 2011;22(1): 27–35.
24. Clark SJ, Houle B. Validation, replication, and sensitivity testing of Heckman-type selection models to adjust estimates of HIV prevalence. *PLoS One.* 2014;9(11):e112563.
25. McCabe SE, West BT. Selective nonresponse bias in population-based survey estimates of drug use behaviors in the United States. *Soc Psychiatry Psychiatr Epidemiol.* 2016; 51(1):141–153.

26. West BT, Little RJA. Nonresponse adjustment of survey estimates based on auxiliary variables subject to error. *J R Stat Soc Ser C Appl Stat*. 2013;62(2):213–231.

27. Andridge RR, Little RJA. Extensions of proxy pattern-mixture analysis for survey nonresponse. *American Statistical Association Proceedings of the Survey Research Methods Section*. 2009;2468–2482.

28. Andridge RR, Little RJA. Proxy pattern-mixture analysis for survey nonresponse. *J Off Stat*. 2011;27(2):153–180.

29. SAS Institute, Inc. *The MI Procedure: Adjusting Imputed Values in Pattern-Mixture Models. SAS/STAT(R) 13.1 User's Guide*. Cary, NC: SAS Institute, Inc.; 2015.

30. SAS Institute, Inc. *The MI Procedure: Adjusting Imputed Classification Levels in Sensitivity Analysis. SAS/STAT(R) 13.1 User's Guide*. Cary, NC: SAS Institute, Inc.; 2015.

31. Peress M. Correcting for survey nonresponse using variable response propensity. *J Am Stat Assoc*. 2010;105(492): 1418–1430.

32. Alanya A, Wolf C, Sotto C. Comparing multiple imputation and propensity-score weighting in unit-nonresponse adjustments: a simulation study. *Public Opin Q*. 2015;79(3): 635–661.

33. Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*. 2009;96(3):723–734.

34. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22(4):523–539.

35. Zhang G, Little R. A comparative study of doubly robust estimators of the mean with missing data. *J Stat Comput Simul*. 2011;81(12):2039–2058.