When Data Are Not Missing at Random: Implications for Measuring Health Conditions

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2011-000696
Article Type:	Research
Date Submitted by the Author:	25-Jan-2012
Complete List of Authors:	Frankel, Martin; Baruch College, CUNY, Statistics Battaglia, Michael; Abt Associates Inc, SSM Division Balluz, Lena; Centers for Disease Control and Prevension, Strine, Tara; Centers for Disease Control and Prevension,
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Epidemiology, Public health, Mental health
Keywords:	MENTAL HEALTH, MISSING DATA, BRFSS, IMPUTATION

SCHOLARONE™ Manuscripts

When data are not missing at random: Implications for measuring health conditions

ARTICLE FOCUS:

1. The article addresses issues associated with the fact that when cross-sectional surveys are used to estimate public health conditions and behaviors some respondents do not answer all of the questions. This is referred to as item nonresponse.
2. While “weighting” is used to address overall non-response, the development of weights for the subset of respondents answering each question is impractical.
3. The tabulation of specific estimates (related to a question) based on persons responding to the question may result in survey bias.
4. A number of imputation techniques have been developed that address the resulting bias associated with the restriction of tabulations to question responders only.

KEY MESSAGES:

1. Restricting survey estimates to overall survey responders only (eliminating question specific non-responders may produce biased survey estimates.
2. Standard methods of question specific imputation may eliminate or reduce some of this bias.
3. Techniques that take advantage of the non-random question specific non-response within standard demographic groups may further reduce this estimation bias.

STRENGHTS AND LIMITATIONS

The article shows that standard methods for item imputation may fall short of maximum possible bias reduction. If additional (non-demographic) correlates of reporting among responders are present, these may be used to improve non-response imputation models. The article is focused on the self-reporting of anxiety and depression levels in Behavioral Risk Factor Surveillance System (BRFSS), a RDD telephone survey. Reports of conditions other than anxiety and depression and in non RDD surveys may not be amenable to this non-response modeling for imputation.

Abstract

Weighting a sample using socio-demographic variables is traditionally used as a potential solution for missing respondent level data. For missing item level data the picture is less consistent. The two most common methods of addressing the issue of missing item information are to do nothing (tabulate only responders to the item) or simple socio-demographic item nonresponse imputation. We show that for certain patterns of missing data, imputation that makes used of additional related variables may be more appropriate in the reduction of item nonresponse bias. In fact, when estimates are based on standard demographic imputation or no imputation at all, the results might be quite misleading. In the collection of anxiety and depression data in the BRFSS we have found that questions related to patterns of missing data exist. In this paper we show methods of imputation that may be used to compensate for this type of question related nonrandom missing data. These methods provided a statistically significant improvement (bias reduction) in the resulting survey estimates.

The use of statistical weights to compensate or adjust for person level (case) nonresponse and non-cooperation has become part of generally accepted practice health surveys. For example, the three largest federally funded health surveys: The National Health Interview Survey (NHIS), The National Health and Nutrition Examination Survey (NHANES) and The Behavioral Risk Factor Surveillance System (BRFSS) all use respondent level weights in order to produce various estimates of health risks and health behaviors (1-3). This consistency in treatment of (person) case level nonresponse non-cooperation is lacking with respect to (item) question specific nonresponse. One of the reasons for this lack of consistency is the fact that there is a diversity of opinion about the use of the imputation on the part of "survey experts." This is also reflected in the imputation literature (4-5).

In this paper we discuss the possible impacts of imputation or the absence of imputation in surveys that are intended to estimate and understand various health conditions and health risks. In particular we show that there are situations where non-imputation or even the use of standard socio-demographic based imputation methods may produce substantial bias in the estimation of certain health conditions and risks.

We compare various estimates of health behavior and risk that result from no-imputation, standard socio-demographic based imputation and finally imputation that is based on the use of all possible covariates in the survey. We show that when there is a moderate degree of association with variables that are missing and other non-missing variables, then the lack of imputation may lead to various degrees of item nonresponse bias.

Materials and Methods

BRFSS Anxiety and Depression Module

The Behavioral Risk Factor Surveillance System (BRFSS) is the largest health survey in the U.S. The BRFSS is conducted annually in each of the 50 states and the District of Columbia by the Centers for Disease Control and Prevention. This state-based survey is conducted by telephone with a sample of adults (age 18+) using random-digit-dialing. The BRFSS questionnaire consist of a core module that collects basic risk factor and health condition data such as general health, health care coverage, smoking, alcohol use, asthma and BMI, as well as socio-demographic characteristics such as age gender race/ethnicity and education. The core section is followed by one or more topic-specific modules. States determine which modules will be administered in a given year. Examples of modules include adult asthma history, anxiety and depression, diabetes and intimate partner violence. The BRFSS weighting methodology involves the calculation of a design weight that accounts for the probability of selection of the adult. The design weight then undergoes poststratification to state level population control totals using age group, gender and race/ethnicity.

In 2006 355,710 interviews were conducted with adults. Our focus is on the 218,726 adults who were administered the anxiety and depression module in 39 states. This module is modeled after the Patient Health Questionnaire 8 (PHQ-8) (6). The first eight questions are PHQ-8 which consists of eight of the nine DSM-IV criteria for diagnosis of major depression.

"Now, I am going to ask you some questions about your mood. When answering these questions, please think about how many days each of the following has occurred in the past 2 weeks."

1. "Over the last 2 weeks, how many days have you had little interest or pleasure in doing things?"

2. "Over the last 2 weeks, how many days have you felt down, depressed or hopeless?"

3. "Over the last 2 weeks, how many days have you had trouble falling asleep or staying asleep or sleeping too much?"
4. "Over the last 2 weeks, how many days have you felt tired or had little energy?"

5. "Over the last 2 weeks, how many days have you had a poor appetite or ate too much?"

6. "Over the last 2 weeks, how many days have you felt bad about yourself – or that you were a failure or had let yourself or your family down?"

7. "Over the last 2 weeks, how many days have you had trouble concentrating on things, such as reading the newspaper or watching TV?"

8. "Over the last 2 weeks, how many days have you moved or spoken so slowly that other people could have noticed? Or the opposite –being so fidgety or restless that you were moving around a lot more than usual?"

A depression severity scale is created by scoring the PHQ-8 by converting the number of days for each question to points:

- 0-1 day = 0 points
- 2-6 days = 1 point
- 7-11 days = 2 points
- 12-14 days = 3 points

The number of points is totaled across the eight questions in order to determine the depressive symptoms severity score:

- 0-4 points = no depression
- 5-9 points = mild depression
- 10-14 points = moderate depression
- 15-19 points = moderately severe depression
- 20+ points = severe depression

If any of the 8 questions are missing, a score is not calculated. Adults with a severity score are then divided into two categories using a score of 10 or higher versus less than 10. A score of 10 or higher has 88% sensitivity and specificity for major depression. We refer to this dichotomous depression measure as DEP10.

One area of major concern for DEP10 is the level of missing data also referred to as item nonresponse. Of the 218,726 adults administered the anxiety and depression module, 26,878 (12.3%) are missing on DEP10. This is considerably higher item nonresponse rate compared to core module questions like education (0.3%) and alcohol use in the past 30 days (1.0%). The higher level of missing data is related to the placement of the anxiety and depression module later in the questionnaire and the requirement that all 8 questions must be answered to calculate DEP10. With the high level of DEP10 item nonresponse, prevalence estimates calculated using the 191,848 adults with a non-missing DEP10 may be subject to item nonresponse bias for all 39 states combined and at the individual state level.

Item Nonresponse Imputation

Imputation for item nonresponse is now widely used in survey research. One aspect common to single and multiple imputation methods is the use of socio-demographic variables in the imputation process (7). We illustrate the imputation of our dichotomous DEP10 variable using logistic regression to derive a single imputed value (predicted probability). Following the usual approach of identifying socio-demographic variables to include as predictor variables in a weighted logistic regression model for the 191,848 adults with a non-missing DEP10 value, the core BRFSS contains the socio-demographic predictors:

- age
- education
- employment status
- household income
- race/ethnicity
- number of adults in household
- marital status
- veteran status
- currently pregnant

The DEP10 dependent variable is coded to 1 if the adult is positive on the depression scale (score of 10 or higher) and 0 if they are negative on the depression scale (score less than 10). Table 1 presents the adjusted odds ratios from the weighted "socio-demographic" imputation model. Adults who are unable to work are 7.1 more times likely than adults who are currently employed for wages to score positive on the depression scale. The area under the receiver operating characteristic (ROC) curve is 0.763 which is considered acceptable discrimination. (0.50 suggests no discrimination) (8). Compared to a value of 0.50, this ROC level is statistically significant with a p value of 0.0000 (9).

The 2006 BRFSS core questionnaire however contains three mental health related variables that were added as predictors to the "socio-demographic" model. The first question relates directly to mental health status: "Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?" Most responses are in the 0-7 day range or 30 days with the remaining responses tending to clump at 10 and 20 days. We therefore created a 10-category predictor using values of 0 days, 1, 2, 3, 4, 5, 6, 7, 8-29 and 30 days. The second questions measures the impact of poor health on usual daily activities: "During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?" We created a 10-category predictor using the same 10 categories as the mental health status variable. The third questions measures life satisfaction: "In general, how satisfied are you with your life?" This question has 4 response categories: very satisfied, satisfied, dissatisfied, and very dissatisfied.

Including these three single-question proxy depression measures in the core BRFSS questionnaire potentially allows us to do a better job reducing item nonresponse bias if the 12.3% of sample adults with a missing DEP10 value differ from the DEP10 item respondents. As a preliminary step we cross-tabulated each of the three variables by a three categories of DEP10. The results in Table 2 shows that adults who are positive on DEP10 are much more likely to have poor mental health, activity limitation and to be dissatisfied with their lives than adults who score negative on

DEP10. Furthermore, adults who are missing on DEP10 are more likely to have poor mental health, poor health and to be dissatisfied with their lives than adults who are non-missing on DEP10.

Adding these three predictors to the logistic regression model yielded the "socio-demographic/mental health" model shown in Table 3. Adults who reported that their mental health was not good for past 30 days were 11.5 times more likely to score positive on the depression scale than adults who reported 0 days. Adults who were dissatisfied with their lives were 11.4 times more likely to score positive on the depression scale than adults who reported being very satisfied. We also find that adults who reported activity limitation for the past 30 days were 4.9 times more likely to score positive on the depression scale than adults who reported 0 days of poor health. The area under the ROC for this model is 0.911 which is considered outstanding discrimination and is a substantial increase over the "socio-demographic" model. Further, this improvement of 0.148 in the ROC level is statistically significant a p value of 0.0000 (9).

The final step in the imputation process involved using the coefficients of the "socio-demographic" model and from the "socio-demographic/mental health" model to assign predicted probabilities on DEP10 for the 26,878 adults who are missing on DEP10.

We also applied a multiple imputation process using 5 imputations for both of these imputation models (10). Using multiple imputation we find that the changes in DEP10 are all statistically significant (the percent difference between no-imputation and socio-demographic only imputation is 2.3 percent with p value of 0.0000 (11). The percent difference between no-imputation and the imputation with socio-demographics and other proxies is 9.2 percent with p value of 0.0000. The percent difference between imputation with socio-demographics and other proxies and socio-demographics only is 6.7 percent. This is significant at the level p= 0.0000.

Validating the Imputation

A final aspect of our analysis of the DEP 0 imputation involved imputing DEP10 for adults who are non-missing on DEP10 and to then compare the imputed value with their actual value. To implement this validation step we divided the 191,848 adults who are non-missing on DEP10, on a state-by-state basis, into two equal-sized random halves: test sample and validation sample. We then fit the "socio-demographic" model and the "socio-demographic/mental health" model on the test sample. The coefficients of each model were then used to calculate DEP10 predicted for the adults in the test sample. For the comparison of the actual DEP10 value with the imputed value using 2 x 2 cross-tabulations, we first used stochastic rounding to convert the predicted probabilities to 0 or 1 values (12). For example, a predicted probability of 0.70 has a 70% chance of being rounded to 1 (positive) and a 30% chance of being rounded to 0 (negative).

Results

For each state and for all 39 states combined we have three DEP10 estimates: 1) prevalence estimate ignoring adults with missing DEP10 values, 2) prevalence estimate with missing DEP10 values imputed using the socio-demographic model, and 3) prevalence estimate with missing DEP10 values imputed using the socio-demographic/mental health model. The three corresponding prevalence estimates, for the 39 states combined, are 8.7%, 8.9%, and 9.5%. Compared to not imputing missing DEP10 values the prevalence estimate based on the socio-demographic/mental health model increased by 9.2 percent. This is considerably larger than the 2.3 percent increase in DEP10 prevalence resulting from imputing missing values using the socio-demographic model. Thus, without the use of additional variables in imputation, we would understate prevalence by close to 10%. At the state level we find that the percent differences are considerably larger for the socio-demographic/mental health model with increases in DEP10 prevalence as large as 22 percent. We also find that all of the state increases in DEP10 prevalence resulting from the socio-demographic/mental health imputation model are statistically significant at the 0.05 level, after making a Bonferroni correction to the p values, except for the 1.2 percent difference in Delaware.

The validation sample results shown in Tables 5 and 6 demonstrate that the socio-demographic/mental health model is able to more accurately classify adults on DEP10. The socio-demographic/mental health imputation model misclassified 10.0% of adults while the socio-demographic model misclassified 14.1% of adults. Examining the margins for the two tables we observe that the socio-demographic/mental health model yields margins that are closer to the actual margins.

Discussion

While our analysis is restricted to a single set of measures and estimates, the results clearly demonstrate that the data missing completely at random and the data missing at random assumptions may not hold for certain health related survey estimates. Further we show that the use of socio-demographic and proxy driven logistic regression (imputation) may result in improved estimates in the sense that they are statically different from estimates derived by excluding missing data. Given that the imputation process is shown to correctly reproduce nearly unbiased marginal estimates among individuals with known response, the assumption of valid marginal results when the imputation is applied to observations with missing data appears to be supported. Further, since there are statistically different estimates obtained when these imputation procedures are applied to persons with missing data the hypothesized improvement in estimation is supported.

We note that both our socio-demographic only and socio-demographic plus related question imputations were derived using the association of these variables with the appropriate outcome measure. We conclude that the statistically different results obtained by the addition of these imputations are bias reduction. More specifically we conclude that the resulting estimates are closer to those that would obtain with a full enumeration of the sample with no missing item level data. We believe that the general strategy item imputation based on socio-demographic measures as well as a systematic search for relationships between the question with missing data and other survey questions with lower levels of item nonresponse should be adopted as part of sound survey research practice.

We focused on the self-reporting of anxiety and depression levels in Behavioral Risk Factor Surveillance System (BRFSS), a RDD telephone survey. Reports of conditions other than anxiety and depression and in non RDD surveys may not be amenable to this non-response modeling for imputation.

References:

1. Botman SL, Moore TF, Moriarity CL, et al. Design and estimation for the National Health Interview Survey, 1995–2004. *National Center for Health Statistics. Vital Health Stat*. 2000;2(130).

2. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). *National Health and Nutrition Examination Survey Data*. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2007-2008, http://www.cdc.gov/nchs/nhanes.htm.

3. Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System operational and user's guide*, Version 3.0. Atlanta, GA: US Department of Health and Human Services, CDC; 2006. ftp://ftp.cdc.gov/pub/Data/Brfss/userguide.pdf

4. Korn EL, Graubard, BI. *Analysis of Health Surveys*. 1999, John Wiley & Sons, New York.

5. Kalton, G. *Compensating for Missing Survey Data*. 1983. Institute for Social Research, Ann Arbor.

6. Spitzer RL, Kroenke K, Williams JB. Validation and Utility of a Self-report Version of PRIME-MD: The PHQ Primary Care Study and the Patient Health Questionnaire Primary Care Study Group. 1999. *Journal of the American Medical Association*, 282:1737-1744

7. Shafer J. *Analysis of incomplete multivariate data*. 1997. Chapman and Hall, London.

8. Hosmer DW, Lemeshow, S. *Applied Logistic Regression*, Second Edition. 2000. John Wiley & Sons, New York.

9. Izrael D, Battaglia AA, Hoaglin DC, et al. SAS Macros and Tools for Working with Weighted Logistic Regression Models That Use Survey Data. 2003,*SAS Users Group International 28 Proceedings*, Paper 257-28.

10. Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. 1987, John Wiley & Sons, New York.

11. Berglund PA. An Introduction to Multiple Imputation of Complex Sampler Data using SAS v9.2. 2010. *SAS Global Forum Proceedings*, Paper 265-2010.

12. Shlomo N. Statistical Disclosure Control Methods for Census Frequency Tables. 2007. *International Statistical Review*. Volume 75, Issue 2, pages 199–217.

Table 1 Adjusted Odds Ratios for the "Socio-demographic" Model

		Adjusted Odds Ratio Estimates		
Predictor	Effect	Point Estimate	95% Wald Confidence Limits	
Age group	25-34 vs. 18-24	1.048	0.984	1.117
	35-44 vs. 18-24	1.036	0.969	1.107
	45-54 vs. 18-24	0.991	0.918	1.070
	55-64 vs. 18-24	0.693	0.635	0.756
	65-74 vs. 18-24	0.376	0.335	0.422

	75-79 vs. 18-24	0.311	0.267	0.362
	80-99 vs. 18-24	0.242	0.205	0.285
Gender	Male vs. Female	0.619	0.586	0.654
Education	Grades 1 - 8 vs. Never attended school	0.569	0.451	0.718
	Some high school vs. Never attended school	0.744	0.592	0.935
	High school graduate vs. Never attended school	0.511	0.407	0.641
	Some college vs. Never attended school	0.489	0.389	0.614
	College graduate vs. Never attended school	0.323	0.257	0.407
Employment status	Self-employed vs. Employed for wages	1.048	0.977	1.125
	Out of work for more than 1 year vs. Employed for wages	3.011	2.761	3.284
	Out of work for less than 1 year vs. Employed for wages	2.419	2.248	2.602
	A homemaker vs. Employed for wages	1.319	1.235	1.408
	A student vs. Employed for wages	0.847	0.774	0.928
	Retired vs. Employed for wages	1.489	1.373	1.615
	Unable to work vs. Employed for wages	7.058	6.668	7.470
Household	$10,000 to less	0.926	0.854	1.005

income	than $15,000 vs. Less than $10,000			
	$15,000 to less than $20,000 vs. Less than $10,000	0.900	0.832	0.973
	$20,000 to less than $25,000 vs. Less than $10,000	0.775	0.717	0.837
	$25,000 to less than $35,000) vs. Less than $10,000	0.718	0.665	0.775
	$35,000 to less than $50,000) vs. Less than $10,000	0.561	0.518	0.607
	$50,000 to less than $75,000) vs. Less than $10,000	0.475	0.437	0.516
	$75,000 or more vs. Less than $10,000	0.337	0.309	0.366
	Income not reported vs. Less than $10,000	0.491	0.454	0.530
Race/ethnicity	Black only, non-Hispanic vs. White only, non-Hispanic	0.755	0.713	0.800
	Asian only, non-Hispanic vs. White only, non-Hispanic	0.387	0.323	0.464
	Native Hawaiian or Pacific Islander only vs. White only, non-Hispanic	1.023	0.796	1.315
	American Indian or Alaskan Native only vs.	1.369	1.206	1.555

	White only, non-Hispanic			
	Other race only, non-Hispanic vs. White only, non-Hispanic	1.278	1.123	1.455
	Multiracial, non-Hispanic vs. White only, non-Hispanic	1.479	1.333	1.641
	Hispanic vs. White only, non-Hispanic	0.669	0.634	0.706
Marital status	Divorced vs. Married	1.586	1.494	1.684
	Widowed vs. Married	1.344	1.229	1.470
	Separated vs. Married	2.657	2.434	2.899
	Never Married vs. Married	1.310	1.238	1.387
	Member of an Unmarried Couple vs. Married	1.553	1.438	1.676
Pregnancy status	Not Pregnant vs. Pregnant	1.035	0.969	1.105
Veteran status	No vs. Yes	0.823	0.773	0.876
Number of adults in household	2 Adults vs. 1 Adult	0.951	0.900	1.005
	3 Adults vs. 1 Adult	1.171	1.102	1.245
	4+ Adults vs. 1 Adult	1.198	1.119	1.283

Table 2. DEP10 response by three core questionnaire proxy depression measures

	DEP10 response			
Core Questionnaire Proxy Measure	Not missing: Negative	Not Missing: Positive	Not Missing	Missing
Mental health not good for past 30	2.8%	30.6%	5.2%	11.7%

days				
Activity limitation past 30 days	2.3%	19.0%	3.7%	8.9%
Dissatisfied or very dissatisfied with life	3.4%	32.4%	5.9%	43.3%

Table 3 Adjusted Odds Ratios for the "Socio-demographic/Mental Health" Model

		Adjusted Odds Ratio Estimates		
Predictor	Effect	Point Estimate	95% Wald Confidence Limits	
Age group	25-34 vs. 18-24	0.822	0.763	0.886
	35-44 vs. 18-24	0.841	0.777	0.911
	45-54 vs. 18-24	0.667	0.607	0.732
	55-64 vs. 18-24	0.543	0.488	0.603
	65-74 vs. 18-24	0.390	0.340	0.448
	75-79 vs. 18-24	0.315	0.263	0.377
	80-99 vs. 18-24	0.248	0.204	0.300
Gender	Male vs. Female	0.611	0.571	0.654
Education	Grades 1 - 8 vs. Never attended school	0.666	0.499	0.890
	Some high school vs. Never attended school	0.842	0.633	1.121
	High school graduate vs. Never attended school	0.607	0.457	0.806
	Some college vs. Never attended school	0.555	0.417	0.738
	College graduate vs. Never attended school	0.439	0.329	0.585
Employment status	Self-employed	0.981	0.903	1.065

	vs. Employed for wages			
	Out of work for more than 1 year vs. Employed for wages	1.508	1.348	1.688
	Out of work for less than 1 year vs. Employed for wages	1.292	1.179	1.415
	A homemaker vs. Employed for wages	1.121	1.037	1.212
	A student vs. Employed for wages	0.754	0.680	0.835
	Retired vs. Employed for wages	1.176	1.066	1.298
	Unable to work vs. Employed for wages	1.871	1.734	2.018
Household income	$10,000 to less than $15,000 vs. Less than $10,000	0.827	0.747	0.917
	$15,000 to less than $20,000 vs. Less than $10,000	0.996	0.904	1.098
	$20,000 to less than $25,000 vs. Less than $10,000	0.865	0.785	0.952
	$25,000 to less than $35,000) vs. Less than $10,000	0.884	0.804	0.971
	$35,000 to less than $50,000) vs. Less than $10,000	0.730	0.663	0.803
	$50,000 to less than $75,000) vs.	0.711	0.644	0.786

	Less than $10,000			
	$75,000 or more vs. Less than $10,000	0.568	0.513	0.628
	Income not reported vs. Less than $10,000	0.748	0.681	0.823
Race/ethnicity	Black only, non-Hispanic vs. White only, non-Hispanic	0.924	0.863	0.989
	Asian only, non-Hispanic vs. White only, non-Hispanic	0.388	0.319	0.471
	Native Hawaiian or Pacific Islander only vs. White only, non-Hispanic	0.763	0.557	1.045
	American Indian or Alaskan Native only vs. White only, non-Hispanic	0.924	0.786	1.087
	Other race only, non-Hispanic vs. White only, non-Hispanic	1.072	0.913	1.260
	Multiracial, non-Hispanic vs. White only, non-Hispanic	1.119	0.986	1.270
	Hispanic vs. White only, non-Hispanic	0.738	0.694	0.786
Marital status	Divorced vs. Married	1.136	1.056	1.222
	Widowed vs. Married	1.048	0.941	1.166
	Separated vs. Married	1.678	1.508	1.868
	Never Married	0.951	0.889	1.017

	vs. Married			
	Member of an Unmarried Couple vs. Married	1.199	1.094	1.313
Pregnancy status	Not Pregnant vs. Pregnant	0.886	0.818	0.960
Veteran status	No vs. Yes	0.802	0.743	0.865
Number of adults in household	2 Adults vs. 1 Adult	1.008	0.943	1.077
	3 Adults vs. 1 Adult	1.242	1.154	1.337
	4+ Adults vs. 1 Adult	1.241	1.143	1.347
Mental health not good	1 day vs. 0 days	1.047	0.905	1.210
	2 days vs. 0 days	1.274	1.146	1.416
	3 days vs. 0 days	1.873	1.683	2.085
	4 days vs. 0 days	3.065	2.696	3.486
	5 days vs. 0 days	2.393	2.185	2.621
	6 days vs. 0 days	2.401	1.924	2.998
	7 days vs. 0 days,	4.302	3.858	4.797
	8-29 days vs. 0 days	6.989	6.603	7.397
	30 days vs. 0 days	11.470	10.766	12.221
Life satisfaction	Satisfied vs. Very satisfied	3.098	2.924	3.282
	Dissatisfied vs. Very satisfied	11.428	10.551	12.379
	Very dissatisfied vs. Very satisfied	9.639	8.706	10.672
Poor health	1 day vs. 0 days	1.160	1.028	1.310
	2 days vs. 0 days	1.369	1.242	1.509
	3 days vs. 0	2.120	1.903	2.361

	days			
	4 days vs. 0 days	1.841	1.585	2.137
	5 days vs. 0 days	2.286	2.071	2.523
	6 days vs. 0 days	1.773	1.388	2.266
	7 days vs. 0 days,	3.482	3.092	3.922
	8-29 days vs. 0 days	4.835	4.541	5.148
	30 days vs. 0 days	4.899	4.544	5.283

Table 4 DEP10 Prevalence Estimates by State and for all 39 States

State	DEP10 estimate ignoring adults with missing values	DEP10 estimate with missing values imputed by "socio-demographic" model	DEP10 estimate with missing values imputed by "socio-demographic/mental health" model	Percent Difference between DEP10 estimate and DEP10 estimate after imputation using "socio-demographic" model	Percent Difference between DEP10 estimate and DEP10 estimate after imputation using "socio-demographic/mental health" model	Difference between DEP10 estimate and DEP10 estimate after imputation using "socio-demographic/mental health" model significant at 0.05 level
Total	8.7	8.9	9.5	2.3	9.2	SIG
Alabama	12.5	12.6	13.5	0.8	8.0	SIG
Alaska	6.7	7.4	8.2	10.4	22.4	SIG
Arkansas	12.2	12.1	12.8	-0.8	4.9	SIG
California	8.8	9.2	9.9	4.5	12.5	SIG
Connecticut	5.8	6.2	6.8	6.9	17.2	SIG
Delaware	8.2	8.1	8.3	-1.2	1.2	NOT SIG
District of Columbia	7.9	8.3	8.8	5.1	11.4	SIG
Florida	8.9	9	9.7	1.1	9.0	SIG
Georgia	8.2	8.6	9.2	4.9	12.2	SIG
Hawaii	7.2	7.3	7.7	1.4	6.9	SIG
Indiana	9.6	9.8	10.3	2.1	7.3	SIG

Iowa	5.8	6.1	6.6	5.2	13.8	SIG
Kansas	6.9	7.2	7.5	4.3	8.7	SIG
Louisiana	9.5	9.9	11.4	4.2	20.0	SIG
Maine	7.4	7.7	8.1	4.1	9.5	SIG
Maryland	7.5	7.5	8.4	0.0	12.0	SIG
Michigan	10.5	10.6	10.9	1.0	3.8	SIG
Minnesota	6.2	6.3	6.4	1.6	3.2	SIG
Mississippi	13	12.9	13.6	-0.8	4.6	SIG
Missouri	9.4	9.5	10	1.1	6.4	SIG
Montana	6.7	7.1	7.5	6.0	11.9	SIG
Nebraska	5.6	5.9	6.3	5.4	12.5	SIG
Nevada	9	9	9.6	0.0	6.7	SIG
New Hampshire	6.8	7.1	7.5	4.4	10.3	SIG
New Mexico	9.3	9.4	9.7	1.1	4.3	SIG
North Dakota	5.3	5.8	6.3	9.4	18.9	SIG
Oklahoma	11.5	11.7	12.5	1.7	8.7	SIG
Oregon	7.5	8	8.4	6.7	12.0	SIG
Rhode Island	8.6	8.7	9.2	1.2	7.0	SIG
South Carolina	8.8	9.2	9.7	4.5	10.2	SIG
Tennessee	10.3	10.5	10.9	1.9	5.8	SIG
Texas	8.5	8.7	9.1	2.4	7.1	SIG
Utah	8.7	8.8	9.1	1.1	4.6	SIG
Vermont	7.1	7.3	7.7	2.8	8.5	SIG
Virginia	7.3	7.6	8.2	4.1	12.3	SIG
Washington	6.4	6.8	7.3	6.2	14.1	SIG
West Virginia	13.7	13.7	14.2	0.0	3.6	SIG
Wisconsin	6.7	7	7.4	4.5	10.4	SIG
Wyoming	7.3	7.6	8.1	4.1	11.0	SIG

Table 5 Validation Sample results for "Socio-demographic" Imputation Model

Validation Sample	"Socio-demographic" model: Imputed DEP10 value		
DEP10 value	Negative	Positive	Total
Negative	84.09	7.21	91.30
Positive	6.92	1.78	8.70
Total	91.01	8.99	100.00

Table 6 Validation Sample results for "Socio-demographic/Mental Health" Imputation Model

Validation Sample	"Socio-demographic/Mental Health" model: Imputed DEP10 value		
DEP10 value	Negative	Positive	Total
Negative	86.27	5.03	91.30
Positive	4.96	3.74	8.70
Total	91.23	8.77	100.00

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	1
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	2
Objectives	3	State specific objectives, including any prespecified hypotheses	2
Methods			
Study design	4	Present key elements of study design early in the paper	2
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	2-4
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	2-4
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	2-4
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	2-4
Bias	9	Describe any efforts to address potential sources of bias	*2-5*
Study size	10	Explain how the study size was arrived at	2, 4
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	2
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	4-6
		(*b*) Describe any methods used to examine subgroups and interactions	4-6
		(*c*) Explain how missing data were addressed	4-6
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	4-6
		(*e*) Describe any sensitivity analyses	5
Results			

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	2, 4
		(b) Give reasons for non-participation at each stage	
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	2
		(b) Indicate number of participants with missing data for each variable of interest	3
Outcome data	15*	Report numbers of outcome events or summary measures	Table 2
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	Table 1, Table 3, Table 4
		(*b*) Report category boundaries when continuous variables were categorized	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Table 5, Table 6
Discussion			
Key results	18	Summarise key results with reference to study objectives	5-6
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	6
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	5-6
Generalisability	21	Discuss the generalisability (external validity) of the study results	6
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Centers for Disease Control and Prevention Contract with Abt Associates

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

When Data Are Not Missing at Random: Implications for Measuring Health Conditions

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2011-000696.R1
Article Type:	Research
Date Submitted by the Author:	07-May-2012
Complete List of Authors:	Frankel, Martin; Baruch College, CUNY, Statistics Battaglia, Michael; Abt Associates Inc, SSM Division Balluz, Lena; Centers for Disease Control and Prevension, Strine, Tara; Centers for Disease Control and Prevension,
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Epidemiology, Public health, Mental health
Keywords:	MENTAL HEALTH, MISSING DATA, BRFSS, IMPUTATION

When data are not missing at random: Implications for measuring health conditions in the Behavioral Risk Factor Surveillance System

ARTICLE FOCUS:

1. The article addresses issues associated with the fact that when cross-sectional surveys are used to estimate public health conditions and behaviors, some respondents do not answer all of the questions. This is referred to as item non-response.
2. While "weighting" is used to address overall (unit) non-response, the development of weights for the subset of respondents answering each question is impractical.
3. The tabulation of specific estimates (related to a question) based on persons responding to the question may result in survey bias.
4. A number of imputation techniques have been developed that address the resulting bias associated with the restriction of tabulations to question responders only.

KEY MESSAGES:

1. Restricting survey estimates to overall survey responders only (eliminating question-specific non-responders) may produce biased survey estimates.
2. Standard methods of question-specific imputation may eliminate or reduce some of this bias.
3. A systematic search among all variables for strong relationships with the target variables for imputation is strongly recommended.

STRENGHTS AND LIMITATIONS

Standard methods for item imputation involving basic demographics may fall short of maximum possible bias reduction. If additional (non-demographic) correlates of reporting among responders are present, these may be used to improve non-response imputation models. The article is focused on the self-reporting of anxiety and depression levels in the Behavioral Risk Factor Surveillance System (BRFSS), a random-digit-dialing telephone survey. Reports of conditions other than anxiety and depression and in non-RDD surveys may not be amenable to this non-response modeling for imputation.

ABSTRACT

Objectives: To examine the effect on estimated levels of health conditions produced from large scale surveys, when either list-wise respondent deletion or standard demographic item-level imputation is employed. To assess the degree to which further bias reduction results from the inclusion of correlated ancillary variables in the item-imputation process.

Design: Large Cross-sectional (US level) household survey.

Participants: 219,726 US adults (18 and over) in the 2006 the Behavioral Risk Factor Surveillance System (BRFSS) Survey. This survey is the largest US telephone survey conducted by the Center for Disease Control and Prevention.

Primary and secondary outcome measures: Estimated rates of severe depression among US adults.

Results: The use of list-wise respondent deletion and/or as well as demographic imputation results in the underestimation of severe depression among adults in the US. List-wise deletion produces underestimates of 9% (8.7% versus 9.5%). Demographic imputation produces underestimates of 7% (8.9% versus 9.5%). Both of these differences are significant at the .05 level.

Conclusions: The use of list-wise deletion and/or demographic only imputation may produce significant distortion in estimating national levels of certain health conditions.

INTRODUCTION AND BACKGROUND

The use of statistical weights to compensate or adjust for person-level (case) non-response has become part of generally accepted practice in health surveys. For example, the three largest U.S. federally funded health surveys, The National Health Interview Survey (NHIS), The National Health and Nutrition Examination Survey (NHANES) and The Behavioral Risk Factor Surveillance System (BRFSS), all use respondent-level weights in order to produce various estimates of health risks and health behaviors.[1-3]. This consistency in treatment of (person) case-level non-response is lacking with respect to (item) question specific non-response. One of the reasons for this lack of consistency is the fact that there is a diversity of opinion about the use of the imputation on the part of "survey experts." This is also reflected in the imputation literature.[4-8]

The use of either implicit or explicit imputation to compensate for item-specific missing data has probably been a part of "practical survey methodology" since the first use of both surveys and censuses. The U.S. Census has made use of explicit item-level imputation since 1940.[9] However, a number of major health surveys such as the BRFSS and the NHIS generally make use of imputation for variables related to the weighting process or a small number of other substantive variables. Many variables associated with health conditions, risks and behaviors do not receive imputed values. Furthermore, basic population estimates derived from the variables are generally based on respondents with non-missing values.[10-11]

The primary purpose of the imputation discussed in this paper is to improve the estimation of simple population percentages. This is similar to the purpose of imputation in the U.S. Census and in a number of health surveys. We note however, that much of the literature on imputation has focused on the use of imputation to improve more complex parameter estimation (e.g., multivariate regression coefficients). In this paper we discuss the possible impacts of imputation or the absence of imputation in surveys that are intended to estimate and understand various health conditions and health risks. In particular we show that there are situations where non-imputation or even the use of standard demographic-based imputation methods may produce substantial bias in the estimation of certain health conditions and risks.

We compare various estimates of health behavior and risk that result from no-imputation, standard demographic-based imputation, and finally imputation that is based on the use of additional covariates in the survey. We show that when there is a moderate degree of association with variables that are missing and other non-missing variables, then the lack of imputation may lead to various degrees of item non-response bias.

In terms of the theoretical framework introduced by Rubin[12] and often cited in the academic literature, missing data may be "Missing Completely at Random" (MCAR) or "Missing at Random" (MAR). Missing Completely at Random is the assumption that there is no dependence on the variable values that are missing with any other variable in the study, including itself. This rather "strong assumption" implies that estimates based on the non-missing values are unbiased estimates of the corresponding population parameters.

The more frequently assumed MAR mechanism is often expressed as Pr(Y missing|Y,X) = Pr(Y missing|X). This means that the conditional probability of missing values of Y, given both variables Y and others X, is equal to the probability associated with missing values of Y and only the other variables X). If the mechanisms that control the missing data process are unrelated to Y and if the data are MAR, then the missing data process is considered "Ignorable"; if not, it is "Non-Ignorable" (i.e., not missing at random).

This framework is quite useful in examining and dealing with missing data, but it should be pointed out that the theory is not, in the strict sense, testable in most real world situations. Most imputation methods assume that missing values are MAR and that by using basic demographic variables as X, it is possible to remove bias due to missing values in the production of basic parameters. This is not surprising, since the assumption that X variables are basic demographics typically determines the choice of variables in basic sample weighting.

In this study we have found that the assumption that X variables are demographic will result in the elimination of some bias, but that further bias reduction results from the inclusion of other variables that are associated with the variable that is subject to higher item non-response.

MATERIALS AND METHODS

BRFSS Anxiety and Depression Module

The Behavioral Risk Factor Surveillance System (BRFSS) is the largest health survey in the U.S. The BRFSS is conducted annually in each of the 50 states and the District of Columbia by the Centers for Disease Control and Prevention.[13] This state-based survey is conducted by telephone with a sample of adults (age 18+) using random-digit-dialing. The BRFSS questionnaire consists of a core module that collects basic risk factor and health condition data such as general health, health care coverage, smoking, alcohol use, asthma and BMI, as well as demographic characteristics such as age, gender, race/ethnicity and education. The core section is followed by one or more topic-specific modules. States determine which modules will be administered in a given year. Examples of modules include adult asthma history, anxiety and depression, diabetes, and intimate partner violence. The BRFSS weighting methodology involves the calculation of a design weight that accounts for the probability of selection of the adult. The design weight then undergoes poststratification to state-level population control totals using age group, gender and race/ethnicity.

In 2006 355,710 BRFSS interviews were conducted with adults 18 and over. Our focus is on the 218,726 adults who were administered the anxiety and depression module in 39 states. This module is modeled after the Patient Health Questionnaire 8 (PHQ-8).[14] The first eight questions are the PHQ-8, which consists of eight of the nine DSM-IV criteria for diagnosis of major depression.

"Now, I am going to ask you some questions about your mood. When answering these questions, please think about how many days each of the following has occurred in the past 2 weeks."

1. "Over the last 2 weeks, how many days have you had little interest or pleasure in doing things?"

2. "Over the last 2 weeks, how many days have you felt down, depressed or hopeless?"

3. "Over the last 2 weeks, how many days have you had trouble falling asleep or staying asleep or sleeping too much?"

4. "Over the last 2 weeks, how many days have you felt tired or had little energy?"

5. "Over the last 2 weeks, how many days have you had a poor appetite or ate too much?"

6. "Over the last 2 weeks, how many days have you felt bad about yourself – or that you were a failure or had let yourself or your family down?"

7. "Over the last 2 weeks, how many days have you had trouble concentrating on things, such as reading the newspaper or watching TV?"

8. "Over the last 2 weeks, how many days have you moved or spoken so slowly that other people could have noticed? Or the opposite –being so fidgety or restless that you were moving around a lot more than usual?"

A depression severity scale is created by scoring the PHQ-8 by converting the number of days for each question to points [14]:

- 0-1 day = 0 points
- 2–6 days = 1 point
- 7–11 days = 2 points
- 12–14 days = 3 points

The number of points is totaled across the eight questions in order to determine the depressive symptoms severity score:

- 0–4 points = no depression
- 5–9 points = mild depression
- 10–14 points = moderate depression
- 15–19 points = moderately severe depression
- 20+ points = severe depression

It is important to note that if any of the 8 questions are missing, a score is not calculated. Adults with a severity score of 10 or higher are classified as severely depressed. This classification of 10 or higher has 88% sensitivity and specificity for severe depression.[14]

One area of major concern for the measure of severe depression is the level of item non-response. Of the 218,726 adults administered the anxiety and depression module, 26,878 (12.3%) are missing on the measure of severe depression, indicating that one or more of the 8 questions was not answered. The levels of item non-response on the 8 questions are similar, ranging from 5.2% on the felt down depressed or hopeless question to 7.3% on the had little interest or pleasure doing things question. A total of 9,174 (4.2%) adults did not answer all 8 questions. This level of item non-response is considerably higher than item non-response in the BRFSS core module for questions like education (0.3%) and alcohol use in the past 30 days (1.0%). The higher level of missing data is primarily related to the placement of the anxiety and depression module later in the questionnaire where interview "breakoffs" are more likely to occur. With the high level of severe depression item non-response, prevalence estimates calculated using the 191,848 adults with a non-missing measure of severe depression may be subject to item non-response bias for all 39 states combined and at the individual state level.

Rather than focusing on the 8 individual questions, the primary interest of the Behavioral Risk Factor Surveillance System was estimation of the proportion of adults with major depression. We therefore focused our efforts on imputing the single severe depression summary measure.

Imputation of Adults with a Missing Measure of Severe Depression

One aspect common to most imputation methods is the use of demographic variables in the imputation process.[15] We illustrate the imputation of our dichotomous measure of severe depression variable using logistic regression to derive a single imputed value. Following the usual approach of identifying demographic variables to include as predictor variables in a weighted logistic regression model for the 191,848 adults with non-missing severe depression, the BRFSS core module contains the following 10 demographic predictors.

- Age group
- Gender
- Education
- Employment status
- Household income
- Race/ethnicity
- Number of adults in household
- Marital status
- Veteran status
- Currently pregnant

The dependent variable for this logistic regression is 1 if the adult is classified as severely depressed (score of 10 or higher) and 0 if score less than 10. The logistic regression model includes demographic and socio-economic variables in the BRFSS questionnaire. We also added a currently pregnant variable because pregnant women may have a different level of anxiety and depression than non-pregnant women. A veteran status indicator variable was also added to the model to account for the effect of military service on anxiety and depression. Examining the logistic regression model we find that all predictors except for currently pregnant are highly significant. For example, adults who are unable to work are 7.1 more times likely than adults who are currently employed for wages to score positive on the depression scale. The R^2 statistic for the demographic model is.0.080.[16] The area under the receiver operating characteristic (ROC) curve is 0.763 which is considered acceptable discrimination. (0.50 suggests no discrimination).[16] Compared to a value of 0.50, this ROC level is statistically significant with a p value of 0.0000.[17] The imputation of severe depression using demographic variables would normally end at this point with the hope or expectation that the demographic model largely eliminated item non-response bias.

The 2006 BRFSS core module however contains three (non-demographic) mental health related variables that were found to be related to both the positive classification of being severely depressed and the level of non-response with respect to one or more of the eight depression score questions. The first question relates directly to mental health status: "Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?" The second questions measures the impact of poor health on usual daily activities: "During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?" The third questions measures life satisfaction: "In general, how satisfied are you with your life?"

Table 1 shows that among persons who answered the 3 core questions and the 8 level of depression questions, the percentage classified as severely depressed was 8.7%. However, when further restricting to respondents who indicated that their mental health was not good for 30 of the past 30 days, the level of severely depressed was 50.6%. Similar high levels of severe depression were found for persons with activity limitations in the past 30 days and those who were dissatisfied with life.

Table 1. Weighted percent classified as severely depressed for total sample and by response to certain BRFSS core questions

Group	Weighted percent classified "severely depressed"

Total sample	8.7%
Yes to: mental health not good in past 30 days	50.6%
Yes to: had activity limitation in past 30 days	44.3%
Dissatisfied or very dissatisfied with life	47.3%

Given the strong relationship with these 3 core questions, which had low rates of non-response ranging from 1.7% to 3.8%, and our outcome measure of severe depression, we next looked at the overall degree to which persons with positive responses to these 3 core questions showed higher rates of non-response to one or more of the eight depression score questions. Table 2 shows that 11.4% of the sample was missing one of the 8 questions required to compute the severe depression classifications; the three core questions had levels of missing data between 22% and 49%. This suggested that the use of these 3 core questions, with their relatively low rates of non-response, should "improve" the imputation process that was based on demographic variables.

Table 2. Weighted rates of non-response to one or more of the 8 depression questions based on all respondents

Group	Score to determine severe depression missing
Total sample	11.4%
Yes to: mental health not good in past 30 days	22.3%
Yes to: had activity limitation in past 30 days	23.6%
Dissatisfied or very dissatisfied with life	48.5%

These variables were added to the demographic model predictors as follows. For the first question, on mental health status, most responses are in the 0–7 day range or 30 days, with the remaining responses tending to clump at 10 and 20 days. We therefore created a 10-category predictor using values of 0, 1, 2, 3, 4, 5, 6, 7, 8–29 and 30 days. For the second question, on the impact of poor health on usual daily activities, we created a 10-category predictor using the same 10 categories as the mental health status variable. The third question, on life satisfaction, has 4 response categories: very satisfied, satisfied, dissatisfied, and very dissatisfied.

Adding these three predictors to the logistic regression model produced what we call the full model. Adults who reported that their mental health was not good for past 30 days were 11.5 times more likely to have severe depression than adults who reported 0 days. Adults who were dissatisfied with their lives were 11.4 times more likely to have severe depression than adults who reported being very satisfied. We also find that adults who reported activity limitation for the past 30 days were 4.9 times more likely to have severe depression than adults who reported 0 days of poor health. The R^2 statistic is 0.210, a considerable improvement over the demographic model. The area under the ROC for this model is 0.911, which is considered outstanding discrimination and is a substantial increase over the demographic model.[16] Further, this improvement of 0.148 in the ROC value is statistically significant a p value of 0.0000.[17]

The final step in the imputation process involved using the coefficients of the demographic model and the full model to assign predicted probabilities between 0 and 1 on the measure of severe depression for the 26,878 adults with missing values. Using the entire sample along with the sampling weights, one can estimate the proportion of adults who are positive on the measure of severe depression. The adults who were not missing on this measure have a value of 1 or 0 while the imputed adults have a value ranging from 0 to 1. Under this scenario the proportion of adults who have severe depression equals the ratio of the sum of the product of the measure of severe depression times the sampling weight to the sum of the sampling weights. One can however obtain almost exactly the same results by first stochastically rounding the imputed value to 1 or 0 before calculating the proportion that are positive on DEP10. The use of stochastic rounding is discussed below.

Our logistic regression approach is a single-imputation technique. We also used multiple imputation with 5 imputations as implemented in SAS PROC MI for both of the imputation models to obtain standard errors for the severe depression prevalence estimates.[18] Following Kish,[19] we accounted for the overlap in the samples being compared in calculating the correct standard error of each difference. For the 39 states combined we find that the differences in severe depression prevalence estimates are all statistically significant. The percentage difference between no-imputation and demographic imputation is 2.3 percent with p value of 0.0000.[20] The percentage difference between no-imputation and the imputation with the full model is 9.2 percent with p value of 0.0000. Most importantly, the percentage difference between imputation with the full model and the demographic model is 6.7 percent. This is significant with a p value of 0.0000.

Validating the Imputation

The "true" validation of an imputation process must logically involve discovering the true values associated with those individuals requiring the imputation itself. For obvious reasons this is generally impossible. However, we felt a "second best, but practical" validation of our process would be to apply the imputation procedure to those individuals for which full severe depression responses were provided. As previously mentioned our imputation model made use of logistic repression followed by "stochastic rounding" of the predicted probabilities.

We also note that for the validation process we were not focused on the correct imputation of severe depression at an individual level, but rather in aggregate. More specifically, for the 39 states combined could we predict the overall proportion of individuals with severe depression?

To implement this validation step we divided the 191,848 adults who are non-missing on the measure of severe depression, on a state-by-state basis, into two equal-sized random halves: test sample and validation sample. We then fit the demographic model and the full model on the test sample. The coefficients of each model were then used to calculate severe depression predicted probabilities for the adults in the test sample. We then used stochastic rounding to independently convert each of the predicted probabilities to a 0 or 1 value.[21] For example, based on the generation of a uniform random number, a predicted probability of 0.70 has a 70% chance of being rounded to 1 (positive) and a 30% chance of being rounded to 0 (negative).

RESULTS

In this section we first show two sets of results. The first set shows the overall estimates of the proportion of adults with severe depression using list-wise deletion (only retaining respondents with complete information), using the demographic imputation model, and then using our full imputation model. The second set in this section shows the results of our validation.

Imputation Results

For each state and for all 39 states combined we have three severe depression prevalence estimates: 1) prevalence estimate ignoring adults with missing values, 2) prevalence estimate with missing values imputed using the demographic model, and 3) prevalence estimate with missing values imputed using the full model (see Table 3). The three corresponding prevalence estimates, for the 39 states combined, are 8.7%, 8.9%, and 9.5%. Compared to not imputing missing severe depression values, the prevalence estimate based on the full model increased by 9.2%. This is considerably larger than the 2.3% increase in severe depression prevalence resulting from imputing missing values using the demographic model. Thus, without the use of the three BRFSS core module variables in imputation, we would understate severe depression prevalence by close to 10%. While in certain surveys a change from 8.7% to 9.5% may not be considered of substantive import, the extrapolation of this change to all U.S. adults implies that an additional 1.5 million adults may be considered severely depressed.

Table 3. Severe depression prevalence estimates by state and for all 39 states combined

State	Severe depression prevalence: No imputation	Severe depression prevalence: Demographic model imputation	Severe depression prevalence: Full model imputation
Total	8.7%	8.9%	9.5%*
Alabama	12.5	12.6	13.5
Alaska	6.7	7.4	8.2
Arkansas	12.2	12.1	12.8*
California	8.8	9.2	9.9*
Connecticut	5.8	6.2	6.8*
Delaware	8.2	8.1	8.3
District of Columbia	7.9	8.3	8.8
Florida	8.9	9.0	9.7*
Georgia	8.2	8.6	9.2*
Hawaii	7.2	7.3	7.7*
Indiana	9.6	9.8	10.3*
Iowa	5.8	6.1	6.6*
Kansas	6.9	7.2	7.5
Louisiana	9.5	9.9	11.4*
Maine	7.4	7.7	8.1
Maryland	7.5	7.5	8.4*
Michigan	10.5	10.6	10.9
Minnesota	6.2	6.3	6.4
Mississippi	13	12.9	13.6*
Missouri	9.4	9.5	10.0*
Montana	6.7	7.1	7.5*
Nebraska	5.6	5.9	6.3
Nevada	9.0	9.0	9.6*
New Hampshire	6.8	7.1	7.5*
New Mexico	9.3	9.4	9.7
North Dakota	5.3	5.8	6.3*
Oklahoma	11.5	11.7	12.5*
Oregon	7.5	8.0	8.4
Rhode Island	8.6	8.7	9.2*
South Carolina	8.8	9.2	9.7*
Tennessee	10.3	10.5	10.9
Texas	8.5	8.7	9.1
Utah	8.7	8.8	9.1
Vermont	7.1	7.3	7.7*

State	Severe depression prevalence: No imputation	Severe depression prevalence: Demographic model imputation	Severe depression prevalence: Full model imputation
Virginia	7.3	7.6	8.2
Washington	6.4	6.8	7.3*
West Virginia	13.7	13.7	14.2*
Wisconsin	6.7	7.0	7.4
Wyoming	7.3	7.6	8.1*

* Difference in severe depression prevalence between the full model and demographic model is statistically significant at the 0.05 Bonferroni-adjusted level.

At the state level we find that the percentage differences are considerably larger for the full model with increases in severe depression prevalence as large as 22%. We also find that 23 (59%) of the state increases in severe depression prevalence, when comparing the full model with the demographic model, are statistically significant at the 0.05 level, after making a Bonferroni correction to the p values.

Validation Results

The validation sample results shown in Table 4 demonstrate the superiority of the full model. Based on the actual severe depression values, 8.70% of the adults in the validation sample are severely depressed. When the demographic model is applied to the validation sample, 8.99% of adults are classified as severely depressed, a 3.3 percent overestimation of severe depression. The full model classifies 8.77% of adults as severely depressed, which is a much smaller 0.8 percent difference.

Table 4. Validation Sample Results

	Actual prevalence	Demographic imputation model prevalence	Full imputation model prevalence
Severely depressed	8.70%	8.99%	8.77%

DISCUSSION

While our analysis is restricted to estimates of the proportion of adults with severe depression, the results clearly demonstrate that the data missing completely at random (MCAR) assumption reflected in the no-imputation results and the data missing at random (MAR) assumption reflected in the standard demographic model results may not hold for certain health-related survey measures. We found that the use of demographic and proxy covariate driven logistic regression imputation appears to result in improved estimates in the sense that they are statistically different from estimates derived by excluding missing data or imputing missing data only using demographic variables.

Given that the full imputation model is shown to correctly reproduce nearly unbiased marginal estimates among individuals with known response, the assumption of valid marginal results when the imputation is applied to observations with missing data appears to be supported. Further, since there are statistically different estimates obtained when this imputation procedure is applied to persons with missing data the hypothesized improvement in estimation over the demographics-only imputation model is also supported.

We note that both our demographic-only and full imputation models were derived using the association of these variables with the appropriate outcome measure. We conclude that the statistically different results obtained by the addition of these imputations are due to bias reduction. More specifically we conclude that the resulting estimates are closer to those that would be obtained with a full enumeration of the sample with no missing item-level data.

We believe that the general strategy of item imputation based on demographic measures as well as a systematic search for relationships between a question with missing data and other survey questions with lower levels of item non-response should be adopted as part of sound survey estimation practice. That is, when certain sequences of questions may be viewed as subject to high item non-response, due to the sensitivity of the questions, difficulty of answering the questions, and/or placement of the questions towards the end of the questionnaire, the questionnaire should be reviewed to see if "potentially correlated" proxy questions are included. If not, consideration should be given to adding at least one proxy question.

With regards to the imputation model, our findings suggest that part of the standard imputation process should involve a systematic search for items that may be correlated with the key response measure.

REFERENCES:

1 Botman SL, Moore TF, Moriarity CL, et al. Design and estimation for the National Health Interview Survey, 1995–2004. *Vital Health Stat 2* 2000;**130**:1–31.

2 Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey data. http://www.cdc.gov/nchs/tutorials/nhanes/SurveyDesign/Weighting/intro.htm (accessed 9 Nov 2011).

3 Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System operational and user's guide, Version 3.0.* Atlanta, GA: Centers for Disease Control and Prevention 2006. ftp://ftp.cdc.gov/pub/Data/Brfss/userguide.pdf (accessed 9 Nov 2011).

4 Korn EL, Graubard, BI. *Analysis of Health Surveys.* New York: John Wiley & Sons 1999.

5 Kalton, G. *Compensating for Missing Survey Data.* Ann Arbor, MI: Institute for Social Research 1983.

6 Lin TH. Missing data imputation in quality-of-life assessment. 2006. *Pharmacoeconomics* 2006;**14**;917–25.

7 Shrive FM, Stuart H, Quan H, et al. Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC Med Res Methodol* 2006;**6**:57. doi:10.1186/1471-2288-6-57.

8. Durrant, GB. Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates. *Int J Soc Res Methodol* 2009;**12**:293–304.

9 Cantwell PJ, Hogan H, Styles KM. Imputation, apportionment and statistical methods in the U.S. Census: issues surrounding Utah v. Evans. *U.S. Census Bureau Research Report Series*, *Statistics 2005-01*. Washington, DC: U.S. Census Bureau, Statistical Research Division 2005. http://www.census.gov/srd/www/byyear.html (accessed 2 May 2012).

10 Schenker N, Raghunathan TE, Chiu P, et al. *multiple imputation of family income and personal earnings in the National Health Interview Survey: methods and examples*. 2010. http://www.cdc.gov/nchs/data/nhis/tecdoc4.pdf (accessed 2 May 2012).

11 Centers for Disease Control and Prevention. *2006 Behavioral Risk Factor Surveillance System summary data quality report*. Atlanta, GA: US Department of Health and Human Services, CDC; 2007. ftp://ftp.cdc.gov/pub/Data/Brfss/2006SummaryDataQualityReport.pdf (accessed 2 May 2012).

12 Rubin DB. Inference and missing data. *Biometrika* 1976;**29**:581–92.

13. Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System: At A Glance 2010*. Atlanta, GA: US Department of Health and Human Services, CDC; 2010. http://www.cdc.gov/chronicdisease/resources/publications/AAG/brfss.htm (accessed 2 May 2012).

14 Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: The PHQ Primary Care Study and the Patient Health Questionnaire Primary Care Study Group. *JAMA* 1999;**282**:1737-1744.

15 Shafer J. *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall 1997.

16 Hosmer DW, Lemeshow, S. *Applied Logistic Regression.* 2nd ed. New York: John Wiley & Sons 2000.

17 Izrael D, Battaglia AA, Hoaglin DC, et al. SAS macros and tools for working with weighted logistic regression models that use survey data. Paper presented at: SAS Users Group International 28 Proceedings, Seattle, WA, 2003;paper 257-28.

18 Rubin, DB. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons 1987.

19. Kish L. *Survey Sampling*. New York: : John Wiley & Sons 1965: section 12.4.

20 Berglund PA. An introduction to multiple imputation of complex sampler data using SAS v9.2. *SAS Global Forum Proceedings* 2010;paper 265-2010.

21 Shlomo N. Statistical disclosure control methods for census frequency tables. *International Statistical Review* 2007;**75**:199–217.

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	1
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	2
Objectives	3	State specific objectives, including any prespecified hypotheses	2
Methods			
Study design	4	Present key elements of study design early in the paper	2
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	2-4
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	2-4
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	2-4
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	2-4
Bias	9	Describe any efforts to address potential sources of bias	*2-5*
Study size	10	Explain how the study size was arrived at	2, 4
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	2
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	4-6
		(*b*) Describe any methods used to examine subgroups and interactions	4-6
		(*c*) Explain how missing data were addressed	4-6
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	4-6
		(*e*) Describe any sensitivity analyses	5
Results			

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	2, 4
		(b) Give reasons for non-participation at each stage	
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	2
		(b) Indicate number of participants with missing data for each variable of interest	3
Outcome data	15*	Report numbers of outcome events or summary measures	Table 2
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	Table 1, Table 3, Table 4
		(*b*) Report category boundaries when continuous variables were categorized	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Table 5, Table 6
Discussion			
Key results	18	Summarise key results with reference to study objectives	5-6
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	6
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	5-6
Generalisability	21	Discuss the generalisability (external validity) of the study results	6
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Centers for Disease Control and Prevention Contract with Abt Associates

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.