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

Martin R Frankel,[1] Michael P Battaglia,[2] Lina Balluz,[3] Tara Strine[3]

To cite: Frankel MR, Battaglia MP, Balluz L, *et al.* When data are not missing at random: implications for measuring health conditions in the Behavioral Risk Factor Surveillance System. *BMJ Open* 2012;**2**:e000696. doi:10.1136/bmjopen-2011-000696

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2011-000696).

Received 25 January 2012
Accepted 29 May 2012

[1] Department of Statistics, Baruch College, CUNY, New York, New York, USA
[2] Abt Associates Inc, Cambridge, Massachusetts, USA
[3] Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Correspondence to
Dr Martin Frankel;
mfrankel14@yahoo.com

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: 218 726 US adults (18 years and older) in the 2006 Behavioral Risk Factor Surveillance System Survey. This survey is the largest US telephone survey conducted by the Centers 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 demographic imputation results in the underestimation of severe depression among adults in the USA. List-wise deletion produces underestimates of 9% (8.7% vs 9.5%). Demographic imputation produces underestimates of 7% (8.9% vs 9.5%). Both of these differences are significant at the 0.05 level.

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

ARTICLE SUMMARY

Article focus

- The article addresses issues associated with the fact that when cross-sectional surveys are used to estimate public health conditions and behaviours, some respondents do not answer all the questions. This is referred to as item non-response.
- While 'weighting' is used to address overall (unit) non-response, the development of weights for the subset of respondents answering each question is impractical.
- The tabulation of specific estimates (related to a question) based on persons responding to the question may result in survey bias.
- 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

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

Strengths and limitations of this study

- 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, a random-digit-dialling telephone survey. Reports of conditions other than anxiety and depression and in non-random-digit-dialling surveys may not be amenable to this non-response modelling for imputation.

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 US federally funded health surveys, the National Health Interview Survey, the National Health and Nutrition Examination Survey and the Behavioral Risk Factor Surveillance System (BRFSS), all use respondent-level weights in order to produce various estimates of health risks and health behaviours.[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 US 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 National Health Interview Survey 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 behaviours 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 US 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 (eg, 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 behaviour 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). MCAR 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 $\Pr(Y \text{ missing}|Y,X) = \Pr(Y \text{ missing}|X)$. 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' (ie, not MAR).

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 BRFSS is the largest health survey in the USA. 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-dialling. The BRFSS questionnaire consists of a core module that collects basic risk factor and health condition data, such as general health, healthcare coverage, smoking, alcohol use, asthma and body mass index, 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 aged 18 years and older. Our focus is on the 218 726 adults who were administered the anxiety and depression module in 39 states. This module is modelled 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 totalled 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 eight 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 eight questions was not answered. The levels of item non-response on the eight 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 9174 (4.2%) adults did not answer all eight 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 eight individual questions, the primary interest of the BRFSS 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 <10. The logistic regression model includes demographic and socioeconomic 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 found that all predictors except for currently pregnant are highly significant. For example, adults who are unable to work are 7.1 times more likely than those 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 with 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 three core questions and the eight levels 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.

Given the strong relationship with these three 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 three 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 eight 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 three core questions, with their relatively low rates of non-response, should 'improve' the imputation process that was based on demographic variables.

Table 1 Weighted per cent classified as severely depressed for total sample and by response to certain Behavioral Risk Factor Surveillance System core questions

Group	Weighted per cent classified 'severely depressed'
Total sample	8.7
Yes to: mental health not good in the past 30 days	50.6
Yes to: had activity limitation in the past 30 days	44.3
Dissatisfied or very dissatisfied with life	47.3

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

Group	Score to determine severe depression missing
Total sample	11.4%
Yes to: mental health not good in the past 30 days	22.3%
Yes to: had activity limitation in the 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 days 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 four 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 the past 30 days were 11.5 times more likely to have severe depression than those who reported 0 days. Adults who were dissatisfied with their lives were 11.4 times more likely to have severe depression than those who reported being very satisfied. We also found that adults who reported activity limitation for the past 30 days were 4.9 times more likely to have severe depression than those 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] Furthermore, this improvement of 0.148 in the ROC value is statistically significant with 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 five imputations as implemented in SAS PROC MI (version 9.2) for both of the imputation models to obtain SEs for the severe depression prevalence estimates.[18] Following Kish,[19] we accounted for the overlap in the samples being compared in calculating the correct SE of each difference. For the 39 states combined, we found that the differences in severe depression prevalence estimates are all statistically significant. The percentage difference between no imputation and demographic imputation is 2.3% with a p value of 0.0000.[20] The percentage difference between no imputation and the imputation with the full model is 9.2% with a p value of 0.0000. Most importantly, the percentage difference between imputation with the full model and the demographic model is 6.7%. 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 that 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 regression 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 with 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 US adults implies that an additional 1.5 million adults may be considered severely depressed.

At the state level, we found that the percentage differences are considerably larger for the full model with increases in severe depression prevalence as large as 22%. We also found 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% overestimation of severe depression. The full model classifies 8.77% of adults as severely depressed, which is a much smaller 0.8% difference.

DISCUSSION

While our analysis is restricted to estimates of the proportion of adults with severe depression, the results clearly demonstrate that the data MCAR assumption reflected in the no imputation results and the data 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

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	13.0	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*
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.

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. Furthermore, since there are statistically different estimates obtained when this imputation procedure is applied to persons with missing data, the hypothesised improvement in estimation over the demographic-only imputation model is also supported.

Table 4 Validation sample results

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

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 regard 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.

Contributors MF and MPB did basic analysis and writing. LB and TS provided data and substantive understanding of the scales and models and outcome measures.

Funding This study was supported by the Centers for Disease Control and Prevention.

Competing interests None.

Ethics approval Ethics approval was provided by the Centers for Disease Control and Prevention Institutional Review Board and various health departments.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data may be obtained from http://www.cdc.gov/brfss (2006 interview data file).

REFERENCES

1. Botman SL, Moore TF, Moriarty 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: imputation for WHOQOL-BREF. *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.
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;282:1737–44.
15. Shafer J. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall, 1997.
16. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd edn. 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. *SAS Users Group International 28 Proceedings*. Cary, NC: SAS Institute Inc, 2003:2paper 75–28.
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 sample data using SAS v9.2. *SAS Global Forum Proceedings*. Cary, NC: SAS Institute Inc. 2010: paper 265–2010.
21. Shlomo N. Statistical disclosure control methods for census frequency tables. *Int Stat Rev* 2007;75:199–217.