Research Article

Statistics in Medicine

Received 3 October 2008, Accepted 26 October 2009 Published online 22 December 2009 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3809

Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey‡

Nathaniel Schenker,[a]*† Trivellore E. Raghunathan[b] and Irina Bondarenko[c]

Common data sources for assessing the health of a population of interest include large-scale surveys based on interviews that often pose questions requiring a self-report, such as, 'Has a doctor or other health professional ever told you that you have ⟨health condition of interest⟩?' or 'What is your ⟨height/weight⟩?' Answers to such questions might not always reflect the true prevalences of health conditions (for example, if a respondent misreports height/weight or does not have access to a doctor or other health professional). Such 'measurement error' in health data could affect inferences about measures of health and health disparities. Drawing on two surveys conducted by the National Center for Health Statistics, this paper describes an imputation-based strategy for using clinical information from an examination-based health survey to improve on analyses of self-reported data in a larger interview-based health survey. Models predicting clinical values from self-reported values and covariates are fitted to data from the National Health and Nutrition Examination Survey (NHANES), which asks self-report questions during an interview component and also obtains clinical measurements during a physical examination component. The fitted models are used to multiply impute clinical values for the National Health Interview Survey (NHIS), a larger survey that obtains data solely via interviews. Illustrations involving hypertension, diabetes, and obesity suggest that estimates of health measures based on the multiply imputed clinical values are different from those based on the NHIS self-reported data alone and have smaller estimated standard errors than those based solely on the NHANES clinical data. The paper discusses the relationship of the methods used in the study to two-phase/two-stage/validation sampling and estimation, along with limitations, practical considerations, and areas for future research. Published in 2009 by John Wiley & Sons, Ltd.

Keywords: diabetes; hypertension; measurement error; missing data; multiple imputation; obesity; propensity score; two-phase sampling

1. Introduction

1.1. Overview of the paper

Large-scale surveys based on interviews are often used for assessing the health of populations. Such surveys provide large, representative samples of the populations of interest, rich sets of covariates, and widely available data for public use. Typically, however, the data on health conditions in large-scale surveys are based on responses to questions such as, 'Has a doctor or other health professional ever told you that you have ⟨health condition of interest⟩?' or 'What is your ⟨height/weight⟩?' Use of such self-reported data could lead to inaccuracies in estimating measures of health (see, e.g. [1–7]). For example, some respondents might not have access to a doctor or other health professional, and thus they might answer 'no' even if they have the condition

[a]*National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Room 3209, Hyattsville, MD 20782, U.S.A.*
[b]*Department of Biostatistics, School of Public Health, and Institute for Social Research, University of Michigan Ann Arbor, MI, U.S.A.*
[c]*Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, U.S.A.*
**Correspondence to: Nathaniel Schenker, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Room 3209, Hyattsville, MD 20782, U.S.A.*
†*E-mail: nschenker@cdc.gov*

in question. Other respondents might misreport their height and/or weight. Such inaccuracies could be viewed as types of measurement or response errors, in the sense that answers to the self-report questions might not always accurately measure the clinical values. Alternatively, they could be viewed as types of missing data, in the sense that some of the actual data of interest (e.g. clinical diagnoses or exact body measurements) are not collected in the survey. Under either viewpoint, an analysis seeking to correct for the inaccuracies would use information that relates the observed (mismeasured) values to the unobserved (correctly measured) values [8].

This paper describes research on methods for imputing clinical values for a large-scale survey that relies only on self-report questions for its information on health conditions. The imputations are created using models that have been fitted to data from a smaller survey that obtains both self-reported and clinical data. Such methods could be particularly useful in the context of public-use data, where it is desirable to allow secondary analysts to use data from the large-scale, interview-based survey alone to conduct their analyses; to address a variety of analysis problems, including those of a multivariate nature; and to adjust for differences between self-report and clinical values without needing to develop specialized estimation procedures for each problem. In such a context, a natural approach is to use imputation. The data-collection agency can expend effort and use in-house knowledge in creating imputations of clinical values, and then the secondary analysts can use standard techniques to analyze the data completed by imputation. Note that if an analyst were to prefer self-report values to clinical values, perhaps because of the notion that self-report values reflect a substantial amount of history, whereas clinical values are sometimes based on measurements made during a single visit, the analyst would be free to ignore the imputed values and just analyze the self-reported data.

An illustrative study is presented where the large-scale survey is the National Health Interview Survey (NHIS) of the National Center for Health Statistics (NCHS), an agency within the U.S. Centers for Disease Control and Prevention, and where the smaller survey is NCHS's National Health and Nutrition Examination Survey (NHANES). The conditions considered are hypertension, diabetes, and obesity. Models that relate clinical values to self-reported values, while accounting for differences between the surveys, are fitted to NHANES data for 1999–2002. The fitted models are used to multiply impute clinical values for persons in the 1999–2002 NHIS. Various analyses involving hypertension, diabetes, and obesity are carried out using the multiply imputed data, and the results are compared with analyses using only the self-reported data from the NHIS. Results from analyses using only the clinical data from the NHANES are also examined, as a standard for comparison, under the assumption that if clinical data had been obtained in the NHIS, the estimates based on clinical data from the two surveys would have been similar (since both surveys are nationally representative).

To illustrate the issue of differences between self-reported data and clinical data, Table I displays estimates of prevalence rates for hypertension, diabetes, and obesity by level of education and by race/ethnicity, based on data from the NHANES for people aged 20 and above. Estimates based on self-reported data as well as estimates based on clinical data are shown. For every combination of condition and subgroup in the table, the estimate based on clinical data is larger than the estimate based on self-reported data.

The idea of using models fitted to one set of data, with information about an item under two types of coding or reporting systems, to impute values for a separate set of data, with information under only one of the systems, has been used in other contexts. An early project of this type concerned the incomparability of industry and occupation coding schemes for narrative job descriptions that were used for public-use samples from the 1970 and 1980 censuses. To make a large public-use database from the 1970 census comparable in terms of coding schemes with public-use files from the 1980 census, models fitted to a separate, smaller 'double-coded' sample (having both 1970 and 1980 codes) from the 1970 census were used to multiply impute 1980 codes for the 1970 public-use data [9, 10]. A more recent project concerned the new standards for race reporting in Federal data collections [11], which allow respondents to report more than one race category for the person in question. To make data from the 2000 census comparable to data collected under the prior (1977) standards, which allowed only single-race reporting, a form of imputation was used based on models fitted to the NHIS, which collects race data under both standards [12–14]. The two projects just outlined are reviewed and contrasted in [15]. Although these two projects and the problem described in the current paper have similar basic structures, they differ substantially with regard to their detailed features as well as the imputation methods used to address them.

Table I. Comparison of NHANES Estimated Prevalence Rates for Persons of Ages 20 Years and Above: Self-Reported (SR) Data versus Clinical (CL) Data.

		Hypertension		Diabetes		Obesity	
Categories		SR	CL	SR	CL	SR	CL
Education	< HS Grad.	29.5	38.1	12.4	16.2	28.2	31.2
	HS Grad.	24.5	31.8	6.3	9.2	29.2	32.2
	> HS Grad.	18.3	24.0	4.2	6.1	22.7	26.8
Race/ Ethnicity	Hispanic	14.1	20.5	8.5	10.6	26.6	29.8
	N.H. Black	30.9	38.7	9.2	12.3	33.7	37.1
	N.H. White	22.3	28.8	5.6	8.2	24.3	28.0

Note: Certain records were excluded from the data for this study, as discussed in Section 2.1.2.

Published in 2009 by John Wiley & Sons, Ltd.
Statist. Med. **2010**, 29 533–545

The setup in this paper is related to estimation based on two-phase or double sampling in survey sampling [16] as well as estimation based on two-stage or validation sampling in biostatistics and epidemiology [17, 18]. In such problems, an inexpensive measure is collected for all units in a large study, whereas a more expensive but more accurate measure is collected for a subsample of the units. The information in both sets of measurements is used to obtain a single estimate from the entire study. Examples of the use of multiple imputation in the context of validation sampling are given in [8, 19]. For the problem considered in this paper, the self-report values are analogous to the inexpensive measure, the NHIS is analogous to the large study, the clinical values are analogous to the more accurate measure, and the NHANES is analogous to the units on which the more accurate measure is obtained.

A major difference between the problem considered here and two-phase/two-stage/validation sampling and estimation is that, because the NHANES units are not a subset of the NHIS units, the NHANES units are not included in analyses of the multiply imputed NHIS data. This difference also distinguishes the problem considered here from traditional problems of imputation for missing data, in the sense that the units on which the data are fully observed (the NHANES) are not included in the post-imputation analysis. The emphasis here is on viewing the multiply imputed NHIS file as a stand-alone infrastructure to improve on analyses of self-reported data.

1.2. Some theoretical results

To provide some theoretical insight into the preceding discussion and the results to be presented later, consider the following simple setup. Let (X_i, Y_i), $i=1,\ldots,n+m$ be independent and identically distributed bivariate normal random variables, which are divided into two samples. Suppose that in sample 1, (X_i, Y_i), $i=1,\ldots,n$ are observed, whereas in sample 2, only X_i, $i=n+1,\ldots,n+m$ are observed because the values of Y are missing. Thus, in the context of this paper, X is analogous to the self-report variable, Y is analogous to the clinical variable, sample 1 is analogous to the NHANES, and sample 2 is analogous to the NHIS. Suppose further that it is desired to make inferences about $\mu_Y = \mathrm{E}(Y)$ from sample 2, and that if Y_i, $i=n+1,\ldots,n+m$ had been observed, μ_Y would have been estimated by $\bar{Y}^{(2)}$ and $\mathrm{V}(\bar{Y}^{(2)})$ would have been estimated by $s_Y^{2(2)}/m$, where $\bar{Y}^{(2)}$ and $s_Y^{2(2)}$ are, respectively, the sample mean and variance of Y in sample 2.

Now suppose that multiple, say D, imputations of the missing Y-values have been created for sample 2 based on the imputation model $Y=\beta_0+\beta_1 X+\varepsilon$, $\varepsilon\sim N(0,\sigma_\varepsilon^2)$ fitted to sample 1 using the steps described in [20]. The result is D completed data sets for sample 2, from which D different versions of the complete-data estimates $\bar{Y}^{(2)}$ and $s_Y^{2(2)}/m$ can be calculated, one version for each of the completed data sets. These D sets of estimates can be combined via the standard combining rules for multiple imputation to make inferences about μ_Y [21, 22].

The following theoretical results can be derived using standard results on linear regression, as given in [23], together with manipulations relating unconditional expectations and variances to conditional expectations and variances, such as those performed in [24]. With an infinite number of imputations (i.e. $D\to\infty$), the multiple-imputation point estimator of μ_Y is

$$\hat{\mu}_Y^{MI}=\hat{\beta}_0^{(1)}+\hat{\beta}_1^{(1)}\bar{X}^{(2)}, \tag{1}$$

where $\hat{\beta}_0^{(1)}$ and $\hat{\beta}_1^{(1)}$ are the least-squares estimates based on sample 1, and $\bar{X}^{(2)}$ denotes the sample mean of X in sample 2. The variance of the point estimator is

$$\mathrm{V}(\hat{\mu}_Y^{MI})=\frac{\sigma_Y^2(1-\rho^2)}{n}+\frac{\rho^2\sigma_Y^2}{m}+O(n^{-2}), \tag{2}$$

where σ_Y^2 is the variance of Y and ρ is the correlation between X and Y. Finally, the expectation of the multiple-imputation variance estimator is

$$\mathrm{E}[\hat{\mathrm{V}}^{MI}(\hat{\mu}_Y^{MI})]=\frac{\sigma_Y^2(1-\rho^2)}{n}+\frac{\rho^2\sigma_Y^2}{m}+\frac{2\sigma_\varepsilon^2}{m}+O(n^{-2}). \tag{3}$$

Expression (1) for the multiple-imputation estimator of the mean has the same form as the two-phase regression estimator of the mean [25]. The difference between the two approaches in this case is that with two-phase estimation, $\bar{X}^{(2)}$ would be replaced by a sample mean calculated using the combined sample of $n+m$ units rather than just the m units in sample 2.

It follows from expression (2) that the relative efficiency of $\bar{Y}^{(1)}$, the sample mean of Y in the first sample, to $\hat{\mu}_Y^{MI}$, the point estimator based on the multiple imputations for sample 2, is $\mathrm{RE}(\bar{Y}^{(1)},\hat{\mu}_Y^{MI})=\mathrm{V}(\hat{\mu}_Y^{MI})/(\sigma_Y^2/n)=(1-\rho^2)+\rho^2 n/m+O(n^{-1})$. Three limiting cases are worth mentioning. First, as m becomes very large relative to n, the main contributor to the relative efficiency is ρ, and $\hat{\mu}_Y^{MI}$ tends to gain efficiency relative to $\bar{Y}^{(1)}$ as ρ increases. Second, as ρ approaches 1, the relative efficiency approaches the ratio of the sample sizes. Finally, as ρ approaches 0, the relative efficiency approaches 1.

Comparison of expressions (2) and (3) shows that the multiple-imputation variance estimator tends to overestimate the actual variance of $\hat{\mu}_Y^{MI}$ by the amount $2\sigma_\varepsilon^2/m$ and that the approximate relative bias of the multiple-imputation variance estimator is $2/[(m/n)+\{\rho^2/(1-\rho^2)\}]$. Thus, the conservatism of the variance estimator decreases as m increases relative to n and/or as ρ increases. When $X\equiv 1$, so that imputation is carried out without any predictors, the comparison of expressions (2) and (3) reduces to a result presented in [26] for that simpler case. These results are consistent with results of other authors showing that the

multiple-imputation variance estimator can be conservative when the analyst does not use some of the information that is used by the imputer (in this example, the values of Y in sample 1). See, for example, [27], where this is referred to as a type of 'uncongeniality'.

1.3. Contents of the remainder of the paper

Section 2 describes the data used in this study and provides definitions of the conditions considered. Section 3 describes the imputation models and methods. Section 4 contains results for estimates of prevalence rates and regression analyses involving hypertension, diabetes, and obesity. Section 5 contains a concluding discussion of limitations of the work presented, practical considerations, and topics for further research.

2. Data used and definitions of conditions

2.1. The NHIS and the NHANES

2.1.1. Overview. The NHIS is a multi-purpose health survey and is the principal source of information about the health of the civilian, non-institutionalized, household population of the United States [28]. It has been conducted continuously since its beginning in 1957. For the years considered in this paper (1999–2002), it surveyed roughly 40 000 households containing roughly 100 000 people each year, in 358 primary sampling units, and it oversampled black and Hispanic persons.

The NHIS contains questions on socio-demographic characteristics, health status, activity limitations, injuries, health insurance coverage, and health care that are asked for all people included in the survey. In addition, one adult and one child are sampled from each family, and further information is obtained on items such as health behaviors and conditions.

The NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States [29]. It began in the early 1960s and has been conducted as a series of surveys focusing on different population groups or health topics. In 1999, the NHANES became a continuous survey. For the years 1999–2002, it examined a nationally representative sample of about 5000 persons in 15 primary sampling units each year, with oversampling of low-income persons, adolescents aged 12–19 years, persons aged 60 years and over, African Americans, and Mexican Americans.

The NHANES is unique in that it combines interviews with physical examinations. The interviews, which are conducted in respondents' homes, include demographic, socioeconomic, dietary, and health-related questions. The examinations, which are performed in mobile examination centers, consist of medical and dental examinations, physiological measurements, and laboratory tests.

2.1.2. Target population and data exclusions. The target population in this study was composed of persons of at least 20 years of age, excluding pregnant women, in 1999–2002. Public-use data from the NHIS and the NHANES were used, and the common variables (some recoded) that were used from the two surveys are listed in Table II.

The NHIS sample size for the target population was initially 122 870 persons. To simplify this study, 17 618 persons were excluded from the research data set if they had one or more missing values on the variables listed in Table II. Among those excluded were 8956 persons with self-reported heights outside of the interval [59 inches, 76 inches] or self-reported weights outside of the interval [99 pounds, 285 pounds], for which the values outside of those intervals were set to 'missing' in the NHIS public-use data for purposes of confidentiality. The NHANES sample size for the target population was initially 8889 persons, out of which 2779 persons were excluded using the same criteria as for the NHIS. (There were 365 persons in the NHANES sample with self-reported heights or weights outside of the intervals mentioned above, and they were excluded for consistency with the NHIS data.) Thus, the final sample sizes for the data used in this research were 6110 for the NHANES and 105 252 for the NHIS.

2.2. Definitions of conditions

A person was classified as having hypertension based on self-reported data if a doctor or other health professional had told the person two or more times that he/she had hypertension or high blood pressure. For clinical data, the classification of hypertension was based on having systolic blood pressure greater than 140 mmHg, having diastolic blood pressure greater than 90 mmHg, or taking medication to control blood pressure.

Self-reported diabetes was based on a person having been told by a doctor or other health professional that he/she had high blood sugar or diabetes not related to pregnancy. The clinical classification was based on having glycohemoglobin (HbA1c) greater than or equal to 7 per cent, having fasting plasma glucose (FPG) greater than or equal to 126 mg/dL, or taking medication for diabetes.

Obesity was defined as having a body mass index (BMI) greater than 30, where BMI = (weight in kg)/(height in meters)^2. In calculating BMI, either self-reported height and weight or measured (clinical) height and weight could be used.

Self-reported conditions were available for persons in both the NHANES and the NHIS, whereas clinical conditions were available only for persons in the NHANES, except as noted in Sections 3.1 and 3.2.1 below.

Published in 2009 by John Wiley & Sons, Ltd.
Statist. Med. **2010**, 29 533–545

Table II. Common variables used from the NHANES and the NHIS.

Variable	Values
Year of data collection	1999–2000
	2001–2002
Age (and age^2)	Continuous
Gender	Male
	Female
Race/Ethnicity	Hispanic
	Non-Hispanic black
	Non-Hispanic white
Marital status	Married or lives with a partner
	Separated/divorced
	Widowed
	Never married
Education	Less than high school graduate
	High school graduate
	More than high school graduate
Income poverty ratio	<1
	[1, 2)
	[2, 3)
	$\geqslant 3$
Born in the U.S.A.	Yes
	No
Talked to an MD within one year	Yes
	No
Have Health Insurance	Yes
	No
Have a place to seek medical care	Yes
	No
Self-reported health status	Poor
	Fair
	Good
	Very good
	Excellent
Physical activity	Unable
	Never
	Some
Smoking status	Never
	Former
	Current
Alcohol use	<1 drink per month
	1–12 drinks per month
	>12 drinks per month
Self-reported hypertension status	Yes
	No
Self-reported diabetes status	Yes
	No
Self-reported height	Continuous
Self-reported weight	Continuous

3. Models and methods used for imputation

The imputation methods for the three conditions, hypertension, diabetes, and obesity, have many common features. Therefore, the methods for hypertension will be described first, and then the methods for diabetes and obesity will be described, with emphasis on how they differ from the methods for hypertension.

3.1. Imputation methods for hypertension

For a sampled person in the NHANES or the NHIS, say person i, let R_i, SR_i, CL_i, and X_i denote a survey membership indicator, self-reported and clinical hypertension statuses, and a vector of covariates, respectively, where $R_i = 1$ for the NHANES and $R_i = 0$ for

Published in 2009 by John Wiley & Sons, Ltd.

Statist. Med. ***2010***, 29 533–545

the NHIS; $SR_i=1$ if person i has self-reported hypertension and $SR_i=0$ otherwise; and $CL_i=1$ if person i has clinical hypertension and $CL_i=0$ otherwise. All the variables are observed except for CL_i in the NHIS. (Actually, a question about taking medicine for high blood pressure was asked in the 1999 NHIS but not thereafter. Although answers to this question could have been used to determine the existence of clinical hypertension for some respondents in the 1999 NHIS, the answers were ignored in this study for simplicity.) The goal here is to obtain multiple draws from the predictive distribution for each person in the NHIS, say $D(CL_i|R_i=0,X_i,SR_i)$. Throughout, in the prediction of clinical values, (R_i,X_i,SR_i) are considered fixed, and the models for the predictive distributions are conditional on these quantities.

3.1.1. Using propensity scores for survey membership to create subgroups for modeling. A predictive distribution for CL in the NHIS could be constructed by fitting a logistic regression model to the NHANES data, with CL as the outcome variable and X and SR as the predictors. Then for each person (i) in the NHIS, values of CL_i could be drawn using the fitted logistic regression and the person's values of X_i and SR_i. However, to the extent that the distributions of covariates X in the NHANES and NHIS differ (due to differences in sample designs, sampling variability, non-response, etc.), a model fitted to the NHANES data might not be as appropriate for the NHIS, especially if the model is subject to any misspecification. Moreover, fitting separate 'local' models in different regions of the covariate space would be expected to provide a better fit than fitting a single model globally. Therefore, the general strategy used here is to create subgroups for which the distributions of covariates are similar across the two surveys and then to fit logistic regression models within the subgroups.

To create the subgroups, the method of propensity score subclassification was used [30, 31]. This method is popular in the context of comparing treatments in an observational study, where it is desired to create subgroups for which the covariate distributions are similar across the treatments. In the current context, the data from the NHANES and the NHIS were combined, and then a logistic regression predicting the propensity to be in one survey versus the other was fitted.

The propensity score model can be expressed as

$$\text{logit}\{\Pr(R_i=1|X_i)\}=X_i^{\mathrm{T}}\alpha. \tag{4}$$

Throughout Section 3, X_i will be used in a general way to denote predictors that include variables listed in Table II and selected interactions, although the specific variables and interactions change somewhat across models. Because self-reported hypertension status was to be included as a predictor in the imputation models, it was not included as a predictor in the propensity score model.

Given the fitted model, let $\hat{P}_i^{(H)}$ (with '(H)' denoting hypertension) be the estimated propensity score for person i, that is, the estimated probability that $R_i=1$. Five subgroups of persons were formed based on the quintiles of the distribution of $\hat{P}_i^{(H)}$ in the combined samples. Within the subgroups, $\hat{P}_i^{(H)}$ was also included as a predictor in the imputation models, as described in Section 3.1.3.

3.1.2. Using summaries of the covariates to create parsimonious imputation models. Within the subgroups based on the estimated propensity scores for survey membership, the goal was to predict clinical hypertension for a given self-reported hypertension, controlling for the covariates. Parsimony was needed, given that the propensity score subclassification reduced the sample sizes for fitting each model.

To create a predictor (in addition to the survey membership propensity score) summarizing the covariates X, with emphasis on how the covariates relate to having hypertension, a logistic regression model predicting the propensity of having self-reported hypertension was fitted to the combined survey data within each of the five subgroups for survey membership propensity. The self-reported hypertension propensity score model within subgroup g ($g=1,\ldots,5$) can be expressed as

$$\text{logit}\{\Pr(SR_i=1|X_i,\ \text{subgroup}_i=g)\}=X_i^{\mathrm{T}}\beta_g. \tag{5}$$

Given the fitted model, let $\hat{Q}_i^{(H)}$ denote the estimated propensity score for self-reported hypertension for person i (i.e. the estimated probability that $SR_i=1$). Within each of the five survey membership propensity subgroups, $\hat{Q}_i^{(H)}$ was included as a predictor in the imputation models. To account for variability in self-reported hypertension not explained by $\hat{Q}_i^{(H)}$, SR_i was included as a predictor as well.

3.1.3. Imputing clinical hypertension status for the NHIS. Within survey membership propensity ($\hat{P}_i^{(H)}$) subgroup g ($g=1,\ldots,5$), clinical hypertension status (CL_i) was multiply imputed for persons in the NHIS based on a logistic regression model with the propensity scores, the self-reported status, and all two-way interactions as predictors:

$$\text{logit}\{\Pr(CL_i=1|\hat{P}_i^{(H)},\hat{Q}_i^{(H)},SR_i,\text{subgroup}_i=g)\}=\gamma_{0g}+\gamma_{1g}\hat{P}_i^{(H)}+\gamma_{2g}\hat{Q}_i^{(H)}+\gamma_{3g}SR_i+\gamma_{4g}\hat{P}_i^{(H)}\times SR_i+\gamma_{5g}\hat{Q}_i^{(H)}\times SR_i+\gamma_{6g}\hat{P}_i^{(H)}\times\hat{Q}_i^{(H)}. \tag{6}$$

To create the multiple imputations, model (6) was fitted to the NHANES data, and then the following steps were repeated independently 10 times (resulting in 10 sets of imputations):

1. Under a flat prior distribution for the parameters in model (6), values for the parameters were drawn from their posterior distribution given the NHANES data.
2. The predicted probability of clinical hypertension was calculated for each person in the NHIS based on model (6), conditional upon the values drawn in step 1.

Published in 2009 by John Wiley & Sons, Ltd.
Statist. Med. **2010**, 29 533–545

3. For each person in the NHIS, the clinical hypertension status was drawn from a Bernoulli distribution, conditional upon the predicted probability from step 2.

The number of imputations was limited to 10 to be consistent with the common practice of creating only a small to moderate number of completed data sets in public-use data for the convenience of analysts. The procedure just described for multiple imputation was carried out using the software package IVEware (http://www.isr.umich.edu/src/smp/ive/; accessed August 7, 2009).

3.2. Imputation methods for diabetes

3.2.1. Separating out persons taking medication for diabetes. Both the NHANES and the NHIS collected data on whether medication was being taken for diabetes. Those who were taking such medication were defined as having clinical diabetes and were separated from those not taking such medication for the purpose of imputation. For those who were not taking medication for diabetes, the imputation procedure was similar to that described for hypertension in Section 3.1. However, because there were some missing values of FPG and HbA1c in the NHANES, a preliminary imputation step was performed for the NHANES.

3.2.2. Imputing missing FPG and HbA1c for NHANES persons not taking medication for diabetes. By design, FPG was only measured for a subset of NHANES persons. In addition, although HbA1c was designed to be measured for all persons, there were some missing values of HbA1c as well. Therefore, as a first step, missing values of FPG and HbA1c were multiply imputed (with 10 sets of imputations) for the NHANES via sequential regression multivariate imputation [32], as implemented in IVEware.

Sequential regression multivariate imputation is an iterative technique that cycles through the variables with missing values (FPG and HbA1c in this case), imputing a set of values for each variable using a regression model with predictors including the most recent imputed values for the other variables. The steps used in imputing the set of values are analogous to those described in Section 3.1.3, in that values for the regression parameters are drawn from their posterior distribution, and then the missing values are drawn conditional on the values drawn for the regression parameters. The iterations in sequential regression multivariate imputation are continued until convergence.

For imputing missing values of FPG and HbA1c, the regression models were as follows: $E\{\log(\mathrm{FPG}_i)|X_i, \mathrm{HbA1c}_i\} = X_i^{\mathrm{T}}\eta + \sqrt{\mathrm{HbA1c}_i} \times \eta^*$ and $E(\sqrt{\mathrm{HbA1c}_i}|X_i, \mathrm{FPG}_i) = X_i^{\mathrm{T}}\lambda + \log(\mathrm{FPG}_i) \times \lambda^*$, where FPG_i and $\mathrm{HbA1c}_i$ denote the values of FPG and HbA1c for person i. (In the expressions for the regression models, the use of $\mathrm{HbA1c}_i$ as a predictor for FPG_i and vice versa is intentionally made explicit.) Among the predictors in X_i was the self-reported diabetes status for person i. The transformations for FPG and HbA1c were chosen based on examining residual plots for the regressions fitted to the complete cases. Jeffreys prior distributions were used for the parameters of the regression models.

For each of the NHANES data sets completed by imputation, the existence of clinical diabetes was defined for each person based on the person's (observed or imputed) values of FPG and HbA1c, as described in Section 2.2.

3.2.3. Imputing missing clinical diabetes statuses for the NHIS. Based on each NHANES data set completed as described in Section 3.2.2, a single set of imputations of clinical diabetes statuses was created for the NHIS persons using a procedure completely analogous to that described for hypertension in Section 3.1. The only change was in model (6), for which only a subset of the interactions was used owing to the low rate of self-reported diabetes, with the subset varying across the subgroups for which model (6) was fitted. Creating a single set of imputations for the NHIS corresponding to each of multiple completed NHANES data sets resulted in multiple imputations (10 sets) for the NHIS.

3.2.4. Recombining persons taking and not taking medication for diabetes. After imputations were created for missing clinical diabetes statuses for persons not taking medication for diabetes, those persons were recombined with the persons taking such medication to create full data sets for analysis.

3.3. Imputation methods for obesity

A feature that distinguishes obesity from the other two conditions considered here is that self-reported obesity is derived from two continuous self-reported variables, height and weight (see Section 2.2), which are available in both surveys. Therefore, models relating clinical height and weight to their self-reported counterparts were built using the NHANES and then used to impute clinical height and weight for the NHIS. Given the imputed values of clinical height and weight, clinical obesity was derived.

The first step in the imputation process was completely analogous to that for hypertension; that is, five subgroups of persons were formed based on the quintiles of the distribution of the estimated propensity score for survey membership ($\hat{P}_i^{(O)}$) in the combined samples (with '(O)' denoting obesity and self-reported height and weight not included as predictors in the propensity score model).

Let CLH_i, CLW_i, SRH_i, and SRW_i denote clinical (i.e. measured during an examination) and self-reported height and weight, respectively, for person i. Within survey membership propensity ($\hat{P}_i^{(O)}$) subgroup g ($g = 1, \ldots, 5$), CLH_i, and CLW_i were multiply imputed (with 10 sets of imputations) for persons in the NHIS via sequential regression multivariate imputation using the following regression models:

$$E\{\log(CLH_i)|\hat{P}_i^{(O)}, SRH_i, SRW_i, CLW_i, \mathrm{subgroup}_i = g\}$$
$$= \theta_{0g} + \theta_{1g}\hat{P}_i^{(O)} + \theta_{2g}\log(SRH_i) + \theta_{3g}\log(SRW_i) + \theta_{4g}\hat{P}_i^{(O)} \times \log(SRH_i) + \theta_{5g}\hat{P}_i^{(O)} \times \log(SRW_i) + \theta_g^*\log(CLW_i)$$

and

$$E\{\log(CLW_i)|\hat{P}_i^{(O)}, SRH_i, SRW_i, CLH_i, \text{subgroup}_i = g\}$$
$$= \xi_{0g} + \xi_{1g}\hat{P}_i^{(O)} + \xi_{2g}\log(SRH_i) + \xi_{3g}\log(SRW_i) + \xi_{4g}\hat{P}_i^{(O)} \times \log(SRH_i) + \xi_{5g}\hat{P}_i^{(O)} \times \log(SRW_i) + \xi_g^* \log(CLH_i).$$

The transformations for height and weight were chosen based on examining residual plots for the regressions fitted to the NHANES data. Jeffreys prior distributions were placed on the regression parameters.

4. Results

Sections 4.1 and 4.2 present results for estimating prevalence rates and logistic regression coefficients, respectively. The estimation procedures applied to each complete data set accounted for the complex sample design features (stratification, clustering, and weighting) of the survey data, with variance estimates calculated using jackknife repeated replication as implemented in IVEware. For the analyses of multiply imputed data, each of the completed data sets was analyzed using the design-based procedures just mentioned, and then the results of the multiple analyses were combined using the standard rules (e.g. [21, 22]).

4.1. Estimating prevalence rates

Table III contains estimated prevalence rates for hypertension, diabetes, and obesity, by education and race/ethnicity, based on the NHIS self-reported data and the NHIS multiply imputed clinical data. For every combination of condition and subgroup in the table, the estimate based on the multiply imputed clinical data is larger than that based on the self-reported data, which is analogous to the results discussed in Section 1 for estimates based on NHANES self-reported versus clinical data. In addition, the NHIS estimates based on multiply imputed clinical data are closer to their NHANES clinical counterparts in Table I than are the NHIS estimates based on self-reported data.

The size of the difference between the NHIS estimate based on multiply imputed clinical data and that based on self-reported data is related somewhat to the subgroup of interest, but this 'subgroup effect' is small compared with the effect of using multiply imputed clinical data rather than self-reported data.

Although the goal of this study is to investigate an approach to creating an NHIS data set with imputations of clinical values, a question might arise as to whether there are gains in precision in analyzing such a data set versus simply analyzing the NHANES clinical data. Table IV displays the ratios of the estimated standard errors of the estimated prevalence rates based on the NHANES clinical data to the corresponding estimated standard errors based on the NHIS multiply imputed clinical data. Given that the NHIS sample is about 17 times as large as the NHANES sample in this study (105 252 persons versus 6110 persons), the ratios of estimated standard errors would be expected to be about $\sqrt{17} = 4.1$ or less, as extra variability is introduced into the NHIS estimates via the multiple imputation. Indeed, this is the case for all of the entries in Table IV.

The results in Table IV suggest that multiply imputing clinical values for a large sample using models derived from a smaller sample with known clinical values can result in more precise estimates than those based on the smaller sample alone. Analogous findings were obtained for the industry and occupation coding problem described in Section 1 [10, 33]. As suggested by the theoretical discussion of relative efficiency in Section 1.2, such gains in precision would be expected when there are strong predictors in the imputation model, such as self-report values as predictors of clinical values. In this study, within the five survey membership propensity score subgroups created for imputation, the sample correlation coefficients between self-reported height (or weight) and clinical height (or weight) in the NHANES were all between 0.93 and 0.98; and the kappa statistics for agreement between self-reported statuses and clinical statuses were between 0.54 and 0.68 for hypertension and between 0.56 and 0.81 for diabetes. In addition, R^2 values for various prediction models used in the imputation process (pseudo-R^2 values for binary

Table III. Comparison of NHIS estimated prevalence rates for persons of ages 20 years and above: self-reported (SR) data versus multiply imputed clinical (MICL) data.

		Hypertension		Diabetes		Obesity	
Categories		SR	MICL	SR	MICL	SR	MICL
Education	< HS Grad.	30.9	39.5	11.1	14.2	25.7	30.1
	HS Grad.	22.9	30.1	6.6	8.8	23.5	28.1
	> HS Grad.	16.5	22.8	4.2	6.5	18.7	23.1
Race/ Ethnicity	Hispanic	14.1	20.8	6.9	9.7	23.2	28.2
	N.H. Black	26.7	35.1	8.8	11.3	29.9	34.8
	N.H. White	20.8	27.6	5.6	7.9	19.8	23.1

Note: Certain records were excluded from the data for this study, as discussed in Section 2.1.2.

Published in 2009 by John Wiley & Sons, Ltd.
Statist. Med. **2010**, 29 533–545

outcomes) were in the range of 0.85–0.95 for continuous variables, 0.5–0.7 for hypertension status, and 0.3–0.5 for diabetes status. (The small pseudo-R^2 values for diabetes status were owing to small sample sizes and low prevalences of the disease.)

Note, however, that the ratios in Table IV should be viewed as only approximate indications of relative precision for the following reasons. First, they are subject to sampling variability. Second, as discussed further in Section 5.3, it is suspected that the imputations created in this study did not fully reflect clustering in the surveys. Finally, as discussed in Section 1.2, for the type of setup considered here, in which the NHANES units are omitted from the post-imputation analysis, variance estimates from the multiply imputed NHIS can be conservative.

4.2. Estimating logistic regression coefficients

Thirteen logistic regression models, similar to those found in the literature (e.g. [34–36]), were fitted using (1) the NHANES self-reported data, (2) the NHANES clinical data, (3) the NHIS self-reported data, and (4) the NHIS multiply imputed clinical data. The models, which are outlined in Table V, had a total of 106 coefficients, excluding the intercepts. To facilitate comparisons, the logistic regression coefficients were standardized by multiplying each coefficient by the standard deviation of the predictor corresponding to it (see [37]).

The upper-left graph in Figure 1 displays the standardized coefficients obtained using the NHANES clinical data plotted against those obtained using the NHANES self-reported data, analogous to the comparison of prevalence rates in Table I. Also included are the least-squares line for the regression of the clinical-data-based estimates on the self-reported-data based estimates and the line of equality. The graph of standardized coefficients does not suggest any systematic difference between the clinical estimates and the self-report estimates, unlike the case of prevalence rates. The upper-right graph compares the standardized coefficients based on self-reported data from the two surveys. Again, no systematic difference is displayed. Thus, the upper-left and upper-right graphs would suggest that the standardized coefficients based on the NHIS multiply imputed clinical data should be close to both those based on the NHIS self-reported data and those based on the NHANES clinical data. This is borne out in the lower-left graph (which is analogous to the comparison of prevalence rates in Table III) and the lower-right graph, respectively, except that there is a hint of slight attenuations in some of the coefficients obtained from the NHIS multiply imputed clinical data.

Table IV. Ratios of estimated standard errors of estimated prevalence rates for persons of ages 20 years and above: (NHANES Clinical) ÷ (NHIS Multiply Imputed Clinical).

Categories		Hypertension	Diabetes	Obesity
Education	< HS Grad.	2.2	3.2	2.4
	HS Grad.	1.9	2.1	3.5
	> HS Grad.	2.7	1.9	4.1
Race/ Ethnicity	Hispanic	3.3	3.9	2.8
	N.H. Black	2.1	3.3	2.9
	N.H. White	2.5	2.2	3.7

Note: Certain records were excluded from the data for this study, as discussed in Section 2.1.2.

Table V. Logistic regression models considered.

Model	Gender	Outcome	Predictors (number of coefficients)
1	Males	Hypertension	Age (2); Race/Ethnicity (2); Education (2); Smoking (2); BMI (3); Age*BMI (6); Race/Ethnicity*BMI (6)
2	Females	Hypertension	Same as 1
3	Both	Diabetes	Gender (1); Age (1); BMI (2)
4	Males	Hypertension	Age (1); Race/Ethnicity (2); Smoking (1); Alcohol consumption (1); Obesity (1); Race/Ethnicity*Obesity (2)
5	Females	Hypertension	Same as 4
6	Males	Diabetes	Same as 4
7	Females	Diabetes	Same as 4
8	Both	Hypertension	Race/Ethnicity (2)
9	Both	Hypertension	Education (2)
10	Both	Hypertension	Race/Ethnicity (2); Education (2); Race/Ethnicity*Education (4)
11	Both	Diabetes	Same as 8
12	Both	Diabetes	Same as 9
13	Both	Diabetes	Same as 10

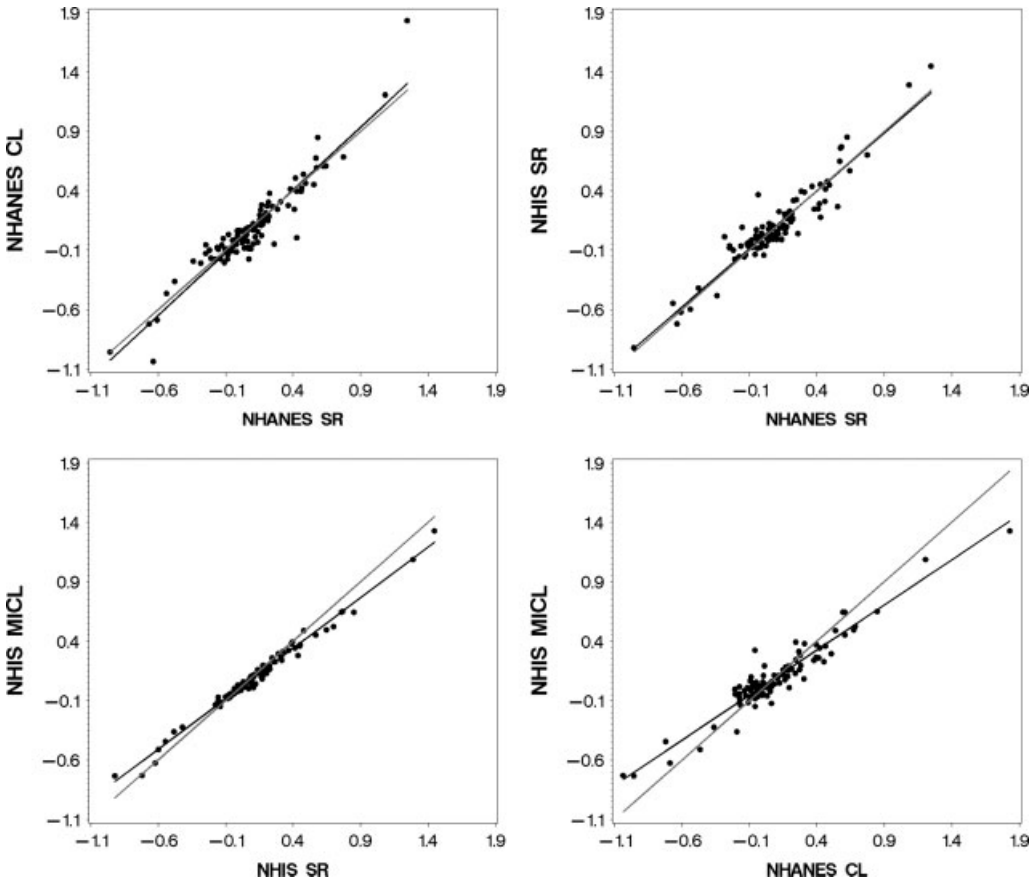

Figure 1. Scatterplots of 106 standardized logistic regression coefficients obtained using (1) the NHANES self-reported (SR) data, (2) the NHANES clinical (CL) data, (3) the NHIS self-reported data, and (4) the NHIS multiply imputed clinical (MICL) data. [Upper left: (2) versus (1). Upper right: (3) versus (1). Lower left: (4) versus (3). Lower right: (4) versus (2). Each graph includes a line of equality (gray) and a least-squares regression line (black).]

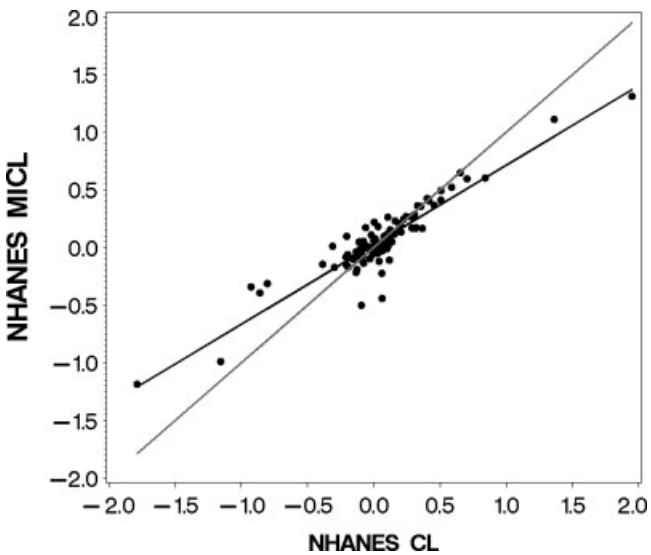

Figure 2. Scatterplot of 106 logistic regression coefficients obtained using NHANES multiply imputed clinical data versus those obtained using the NHANES actual clinical data. [The graph includes a line of equality (gray) and a least-squares regression line (black).]

One possible explanation for the apparent slight attenuations in some of the coefficients based on the NHIS multiply imputed clinical data is that the relations between the outcomes and the predictors in the logistic regression models of Table V were not fully reflected in the imputation models. It is a standard advice that all variables to be used in analyses of imputed data should

Published in 2009 by John Wiley & Sons, Ltd.
Statist. Med. ***2010***, 29 533–545

also be included in the imputation model, to avoid attenuation (see, e.g. [38]). Recall that many predictors entered the imputation models only through the estimated propensity scores. Thus, although the use of propensity score modeling and subgrouping appears to have been mostly effective, the effort to use parsimonious models might have resulted in slight attenuations. Such apparent attenuations also occur when the clinical data in the NHANES are treated as missing and multiple imputations are created for them. See Figure 2, which compares NHANES estimates based on the multiply imputed clinical data to those based on the actual clinical data.

As was the case with prevalence rates (Section 4.1), multiply imputing clinical values for a large sample using models derived from a smaller sample with known clinical values appears to result in more precise estimates than those based on the smaller sample alone. The ratios of the estimated standard errors of the estimated coefficients based on the NHANES clinical data to the corresponding estimated standard errors based on the NHIS multiply imputed clinical data have the following five-number summary: minimum = 1.39, lower quartile = 2.41, median = 3.08, upper quartile = 3.84, and maximum = 11.05. Moreover, a large majority (88 per cent) of the ratios are less than or equal to 4.1, as would be expected given the relative sizes of the samples from the two surveys (see Section 4.1). These results on precision should be interpreted subject to the caveats given regarding the comparisons of precision at the end of Section 4.1.

5. Limitations, practical considerations, and areas for future research

5.1. Assumption of portability

An issue that typically arises when information is combined across data sources is portability, that is, whether a model fitted to one of the data sets applies to the other(s). Section 3.1.1 discussed how subclassification based on the propensity score for survey membership was used to enhance portability by creating subgroups with similar covariate distributions across the NHANES and the NHIS.

Differences in context between the two surveys could also be a source of partial lack of portability. Specifically, although the self-report questions about conditions are comparable across the surveys, answers to self-report questions in the NHANES could differ systematically from those given in the NHIS because the respondents in the NHANES know that they will be examined subsequently. This could be more of an issue for items that might be somewhat sensitive, such as height and weight. (Note that in Tables I and III, the NHANES self-report obesity rates are always a few percentage points higher than the corresponding NHIS self-report obesity rates, whereas the self-report hypertension and diabetes rates are more similar between the surveys.) If it were the case that self-reported values are closer to clinical values in the NHANES than they are in the NHIS, then the procedures discussed in this paper might lead to conservative adjustments for self-reporting in the NHIS, and results based on the multiply imputed values in the NHIS might not be expected to perfectly match those based on the clinical values in the NHANES. Supplemental studies to address such portability issues would be a useful area for future research.

5.2. Issues of uncongeniality

The term 'uncongeniality' [27] refers to situations in which the model used for imputation is incompatible in various ways with the model used for analyzing the data completed by imputation. One type of uncongeniality, owing to not including the data used in fitting the imputation model (the NHANES in this study) in the final data set to be used by analysts (the NHIS in this study), was discussed earlier (see, e.g. Section 1.2).

A second type of uncongeniality could arise when final analyses of the completed data include variables that were not included in the imputation model; this could occur especially if imputations were created for a rich, large-scale survey such as the NHIS. If such omitted variables actually have sizable effects in the final analyses, leaving them out of the imputation model could result in attenuation of the effects. A possible example of such attenuation was discussed in the evaluation of estimated logistic regression coefficients in Section 4.2. If, on the other hand, the omitted variables are 'irrelevant' in the final analysis, then results in [27] suggest that leaving them out of the imputation model will not bias point estimators, but will result in conservative multiple-imputation inferences. Two implications of the omitted-variables issue are that (a) as many relevant variables as possible should be incorporated into the imputation model and (b) documentation of the imputed data should make clear which variables were included in the imputation model, so that subsequent analysts can take the information into account.

5.3. Incorporating complex sample design features in the imputation models

The standard advice for multiple imputation (e.g. [38, 39]) is to incorporate complex sample design features (stratification, clustering, weighting) into imputation models to the extent possible. In the type of problem discussed in this paper, however, it is difficult to fully include the complex sample design features in the imputation models because the sample designs differ across the two surveys. Thus, for example, if survey weights for the NHANES were included as predictors in an imputation model, those weights would not have meaning as predictors if the model were used to create imputations for the NHIS, because the weighting schemes differ across the two surveys. Nevertheless, to the extent that common variables from the two surveys are related to the features of the two sample designs, inclusion of those variables in the imputation models, as was done in this study via the use of propensity scores, will help to reflect the sample designs in the imputation models.

The imputation models used in this study did not include contextual variables such as county-level characteristics in the imputation models, which would have likely helped to reflect clustering in the two surveys further. If the two surveys had the

Published in 2009 by John Wiley & Sons, Ltd.
Statist. Med. **2010**, 29 533–545

same clusters, some other possibilities for reflecting the clustering would be to include indicators or random effects for the clusters in the imputation models; or to implement the imputation process separately within the clusters. The three approaches just discussed (incorporating contextual variables, including cluster indicators or random effects, or imputing within clusters) would require detailed geographical information for the surveys. Such information was not available in this study, because the study used public-use files for the NHIS and the NHANES, which do not include detailed geographical information for reasons of confidentiality. In an actual application by a data producer, however, detailed geographical information would be available.

5.4. Possible refinements of modeling

Several refinements to the modeling used in this study are possible, and they would be worth further research. For example, although the methods used for imputation (Section 3) incorporated features of 'less parametric' approaches such as hot-deck imputation by first creating subgroups based on the estimated propensity scores for survey membership, the procedures within subgroups could be made less parametric as well via use of semi-parametric models (e.g. spline functions or generalized additive models). Another example of a refinement would be to use an approach that borrows strength across the propensity score subgroups. This could result in increased efficiency, and the efficiency gains could allow the incorporation of additional predictors, which could mitigate possible attenuation such as that discussed in Sections 4.2 and 5.2.

5.5. Application with incomplete data on predictors

As mentioned in Section 2.1.2, cases with missing data on any of the common variables listed in Table II were excluded from this study for simplicity. In an actual application, however, it would be feasible to include such cases. For example, the missing values on the common variables could be imputed prior to the implementation of the process for imputing clinical values in the large-scale interview survey (the NHIS in this study). This would be analogous to the imputation of missing FPG and HbA1c values in the NHANES that was described in Section 3.2.2.

Acknowledgements

The authors thank the two referees for comments that improved the paper. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention. A brief overview of the research project described in the paper is given in [40].

References

1. Hamilton LC. Sex differences in self-report errors: a note of caution. *Journal of Educational Measurement* 1981; **18**:221–229.
2. Rowland ML. Self-reported weight and height. *American Journal of Clinical Nutrition* 1990; **52**:1125–1133.
3. Newell SA, Girgis A, Sanson-Fisher RW, Savolainen NJ. The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population: a critical review. *American Journal of Preventive Medicine* 1999; **17**:211–229. DOI: 10.1016/S0749-3797(99)00069-0.
4. Boudreau DM, Daling JR, Malone KE, Gardner JS, Blough DK, Heckbert SR. A validation study of patient interview data and pharmacy records for antihypertensive, statin, and antidepressant medication use among older women. *American Journal of Epidemiology* 2004; **159**:308–317. DOI: 10.1093/aje/kwh038.
5. Scranton RE, Sesso HD, Glynn RJ, Levenson JW, Stedman M, Gagnon D, Gaziano JM. Characteristics associated with differences in reported versus measured total cholesterol among male physicians. *The Journal of Primary Prevention* 2005; **26**:51–61. DOI: 10.1007/s10935-004-0991-z.
6. Ezzati M, Martin H, Skjold S, Vander Hoorn S, Murray CJL. Trends in national and state-level obesity in the USA after correction for self-report bias: analysis of health surveys. *Journal of the Royal Society of Medicine* 2006; **99**:250–257.
7. Raghunathan TE. Combining information from multiple surveys for assessing health disparities. *Allgemeines Statistisches Archive* 2006; **90**:515–526. DOI: 10.1007/s10182-006-0003-0.
8. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* 2006; **35**: 1074–1081 [Commentary: 1081]. DOI: 10.1093/ije/dyl097.
9. Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* 1991; **86**:68–78.
10. Schenker N, Treiman DJ, Weidman L. Analyses of public-use decennial census data with multiply-imputed industry and occupation codes. *Journal of the Royal Statistical Society, Series C* (*Applied Statistics*) 1993; **42**:545–556.
11. Office of Management and Budget. Revisions to the standards for the classification of federal data on race and ethnicity. *Federal Register* 1997; **62**:58781–58790.
12. Ingram DD, Parker JD, Schenker N, Weed JA, Hamilton B, Arias E, Madans JH. United States Census 2000 Population with Bridged Race Categories. *Vital and Health Statistics* 2003; **2**(135):1–55.
13. Schenker N. Assessing variability due to race bridging: application to census counts and vital rates for the year 2000. *Journal of the American Statistical Association* 2003; **98**:818–828. DOI: 10.1198/016214503000000756.
14. Parker JD, Schenker N, Ingram DD, Weed JA, Heck KE, Madans JH. Bridging between two standards for collecting information on race and ethnicity: an application to Census 2000 and vital rates. *Public Health Reports* 2004; **119**:192–205.
15. Schenker N. Bridging across changes in classification systems. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Gelman A, Meng X-L (eds). Wiley: Chichester, 2004; 117–128.

16. Cochran WG. *Sampling Techniques*, Chapter 12 (3rd edn). Wiley: New York, 2007.
17. Breslow NE, Cain KC. Logistic regression for two-stage case–control data. *Biometrika* 1988; **75**:11–20.
18. Pepe MS. Inference using surrogate outcome data and a validation sample. *Biometrika* 1992; **79**:355–365.
19. Yucel RM, Zaslavsky AM. Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association* 2005; **100**:1123–1132. DOI: 10.1198/016214505000000754.
20. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*, Example 5.1. Wiley: New York, 1987.
21. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*, Section 3.1. Wiley: New York, 1987.
22. Harel O, Zhou X-H. Multiple imputation: review of theory, implementation and software. *Statistics in Medicine* 2007; **26**:3057–3077. DOI: 10.1002/sim.2787.
23. Draper NR, Smith H. *Applied Regression Analysis*, Section 1.4 (2nd edn). Wiley: New York, 1981.
24. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* 1986; **81**:366–374.
25. Cochran WG. *Sampling Techniques*, Section 12.6 (3rd edn). Wiley: New York, 1977.
26. Rubin DB, Schenker N. Interval estimation from multiply-imputed data: a case study using agriculture industry codes. *Journal of Official Statistics* 1987; **3**:375–387.
27. Meng X-L. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**:538–558 (Discussion: 558–573).
28. National Center for Health Statistics. 2008 National Health Interview Survey (NHIS) public use data release: NHIS survey description. Division of Health Interview Statistics, National Center for Health Statistics, Centers for Disease Control and Prevention, U.S. Department of Health and Human Services 2009. Available from http://www.cdc.gov/nchs/nhis.htm [accessed August 7, 2009].
29. National Center for Health Statistics. National Health and Nutrition Examination Survey, 2007–2008: overview. National Center for Health Statistics, Centers for Disease Control and Prevention, U.S. Department of Health and Human Services 2007. Available from http://www.cdc.gov/nchs/nhanes.htm [accessed August 7, 2009].
30. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
31. D'Agostino Jr RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998; **17**:2265–2281. DOI: 10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B.
32. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001; **27**:85–95.
33. Treiman DJ, Bielby WT, Cheng M. Evaluating a multiple-imputation method for recalibrating 1970 U.S. census detailed industry codes to the 1980 standard. *Sociological Methodology* 1988; **18**:309–345.
34. Brown CD, Higgins M, Donato KA, Rohde FC, Garrison R, Obarzanek E, Ernst ND, Horan M. Body mass index and the prevalence of hypertension and dyslipidemia. *Obesity Research* 2000; **8**:605–619.
35. Okosun IS, Chandra KMD, Choi S, Christman J, Dever GEA, Prewitt TE. Hypertension and Type 2 diabetes comorbidity in adults in the United States: risk of overall and regional adiposity. *Obesity Research* 2001; **9**:1–9.
36. Gregg EW, Cheng YJ, Cadwell BL, Imperatore G, Williams DE, Flegal KM, Narayan KMV, Williamson DF. Secular trends in cardiovascular disease risk factors according to body mass index in US adults. *Journal of the American Medical Association* 2005; **293**:1868–1874. DOI: 10.1001/jama.293.15.1868.
37. Agresti A. *An Introduction to Categorical Data Analysis*. Wiley: Chichester, 1996; 129.
38. Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**:473–489 (Package: 473–520).
39. Reiter JP, Raghunathan TE, Kinney SK. The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology* 2006; **32**:143–149.
40. Schenker N, Raghunathan TE. Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine* 2007; **26**:1802–1822. DOI: 10.1002/sim.2801.