Multiple Imputation Models and Procedures for NHANES III

Prepared for:

National Center for Health Statistics
Hyattsville, Maryland

Prepared by:

Joseph L. Schafer

June, 2001

Author's academic affiliation: Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802. Direct comments and queries to `jls@stat.psu.edu`.

Summary

This document describes the statistical models and computational methods used to create multiple imputations in the NHANES III Multiply Imputed Data Set. The 33,994 interviewed persons in NHANES III were divided among nine age classes, and a multivariate linear mixed model was applied to each class. These models were designed to reflect interrelationships among variables and key features of the NHANES III sample design. Response variables consisted of select items from the NHANES III examination and the household family, adult, and youth questionnaires for which missing values were to be imputed. Model covariates included demographic descriptors and additional items from the household questionnaires. Transformations were applied when necessary to make distributional assumptions more plausible. Random effects in each model reflected correlations among individuals within primary sampling units, so that the imputations would be compatible with the variance-estimation procedures for complex surveys recommended for analyses of NHANES III. Five sets of multiple imputations were created by Markov chain Monte Carlo procedures. Exploratory and graphical comparisons among observed and imputed values show that important features of marginal distributions and relationships were successfully preserved.

Methods for analyzing the NHANES III Multiply Imputed Data Set are described in the companion technical report, "Analyzing the NHANES III Multiply Imputed Data Set: Methods and Examples." That report also compares results from multiple imputation to those obtained from weighting adjustments for non-examined persons used in previously released NHANES III data sets (DHHS, CD-ROM, Series 11, Number 1A, 1997; Number 2A, 1998).

1 Introduction

In the third National Health and Nutrition Examination Survey (NHANES III), moderate amounts of data became missing due to nonresponse at various stages of the data collection process. Of the 39,695 individuals selected into the NHANES III sample, household interviews were obtained for 33,994 (86%). Among these interviewed persons, 30,818 (91%) were subsequently examined in a Mobile Examination Center (MEC) and 493 (1.4%) received limited physical examinations at home. Rates of response varied across demographic subgroups. Response rates tended to be higher among racial and ethnic (African- and Mexican-American) minorities, persons from larger households, and younger persons. Without corrective measures, estimates from the survey would be biased toward the characteristics of those groups with higher rates of response.

Previously released public-use data sets from NHANES III (DHHS, CD-ROM, Series 11, Number 1A, 1997; Number 2A, 1998) provide sample weights that reflect two stages of adjustment for unit nonresponse. In the first stage, the non-interviewed persons were removed from the sample, and their weights were distributed among interviewed persons with similar demographic characteristics. The resulting adjusted weight (variable `WTPFQX6`) has been recommended for analyses involving items from the household questionnaires. In the second stage, persons who were interviewed but not examined were assigned weights of zero, and their former weights were distributed among examined persons with similar characteristics. The second adjusted weight (variable `WTPFEX6`) has been recommended for analyses involving items from the examination or joint analyses involving household questionnaire and examination items. Details and further guidelines for analysis were provided in *Plan and Operation of the Third National Health and Nutrition Examination Survey*, 1988–94, (NCHS, 1994; DHHS, 1996), and from *NHANES III Reference Manuals and Reports* (DHHS, CD-ROM, 1996).

In addition to nonresponse from non-interviewed and non-examined persons, sporadic missing values occurred on many questionnaire and examination items due to 'Don't know' responses, refusals to answer questions or to submit to examination procedures, examinations that had to be terminated because the subject had to leave early, and so on. For the most

part, no statistical procedures or adjustments were applied to these types of item nonresponse. As a result, users of previously released NHANES III data files will find that many variables include codes for "Blank but applicable," "Don't know," and other instances of failure to obtain usable data.

In 1992, a group of statisticians began to investigate methods of multiple imputation (MI) (Rubin, 1987) to compensate for unit and item nonresponse in NHANES III. This feasibility study culminated with the production of The NHANES III Multiply Imputed Data Set. In MI, each missing value is replaced by several plausible simulated values randomly generated under a statistical model. Each of the several completed data files is analyzed in the same fashion as if it contained no missing values. The several sets of estimates, which randomly vary as a reflection of missing-data uncertainty, are then combined using simple arithmetic to yield a final set of estimates and standard errors. MI can be attractive both to data users and to the data-collecting agency. MI produces 'clean' data files which are easy to analyze. Releasing imputed files helps to ensure that a variety of users performing similar statistical analyses will be led to similar results; variation due to different treatments of missing values by users is removed. Imputation can also be more effective than reweighting in making use of inter-variable relationships to predict missing data values, leading to more efficient estimates (Little, 1986).

After an initial feasibility study (Schafer, Khare, and Ezzati-Rice, 1993), the MI research group concluded concluded that MI offered significant advantages over reweighting in adjusting for nonresponse at the MEC examination stage; substantial gains in precision could result by imputing examination variables for persons who were interviewed but not examined. MI also appeared valuable for handling the sporadic missing values on interview and examination items. However, MI seemed to offer little advantage over reweighting for those who were neither interviewed nor examined. In 1994 and 1995, the research group designed and implemented a simulation study to evaluate the performance of the MI procedure over repeated samples in an NHANES-style survey. This study demonstrated the effectiveness of the method for statistical inferences about means, prevalences, medians, quantiles, and regression coefficients. Details of the simulation procedures and discussion of results are

provided by Little et al. (1995).

Based on these encouraging results, the research group proceeded to develop and implement an MI procedure for NHANES III data as they became available in 1996 and 1997. A set of 67 key variables was designated for imputation, including body measurements, key variables from bone densitometry, fundus photography, blood pressure, and laboratory results from the analysis of blood and urine samples. Five versions of the complete data were produced for all 33,994 interviewed persons. The imputed variables are listed in Table 1, along with the names of the non-imputed variables in previously released NHANES III files to which they correspond. As shown in this table, most of the variables apply to subsets of the sample defined by age. For example, bone densitometry was performed on those 20 and over, whereas fundus photography applied only to those 40 and over.

To simplify the task of modeling and imputation, the sample was split into nine age classes, and a multivariate statistical model was constructed for each class. These models were designed to capture important relationships among these variables and their relationships to other key variables from the home interview: health status, physical activity status, tobacco and alcohol use, self-reported height and weight, home blood pressure readings, and presence of select medical conditions. The models also incorporated basic demographic and economic characteristics of sampled persons and important features of the NHANES III sample design. After five imputations were generated for each age class, the classes were merged back together into five data files. The nine age classes used for imputation and the number of interviewed persons in each are shown in Table 2.

The remaining sections of this document provide technical details of the statistical models and computational algorithms used to create imputations in each age class. Some informal exploratory and graphical comparisons among observed and imputed values are provided to show that the imputation procedures were successful in preserving important features of marginal distributions and inter-variable relationships. Finally, some comparisons are made between estimates and confidence intervals from the NHANES III Multiply Imputed Data Set and those from previously released NHANES III data files.

Table 1: Variables selected for imputation in the NHANES III Multiply Imputed Data Set, and the non-imputed variables in previously released NHANES III public-use files to which they correspond

MI name	Previously	Description	Age range
HOUSEHOLD FAMILY QUESTIONNAIRE ITEMS			
DMPPIRMI	DMPPIR	Poverty income ratio	2 mo +
HFF1MI	HFF1	Anyone living here smoke cigs in home	2 mo +
HOUSEHOLD ADULT QUESTIONNAIRE ITEMS			
HAB1MI	HAB1	Self-rating of health status	17 yr +
HAM5MI	HAM5	How tall are you without shoes-inchs	17 yr +
HAM6MI	HAM6	How much do you weigh in pounds	17 yr +
HAN6SRMI	********	Beer/wine/liquor (recode)	17 yr +
HAQ1MI	HAQ1	Condition of SPS natural teeth	17 yr +
HAR3RMI	********	Smoke cigarettes now (recode)	17 yr +
HAT28MI	HAT28	Compare own activity level to others	17 yr +
HAZAK1MI	HAZA8AK1	K1 for first BP measurement (home)	17 yr +
HAZAK5MI	HAZA8AK5	K5 for first BP measurement (home)	17 yr +
HAZBK1MI	HAZA8BK1	K1 for second BP measurement (home)	17 yr +
HAZBK5MI	HAZA8BK5	K5 for second BP measurement (home)	17 yr +
HAZCK1MI	HAZA8CK1	K1 for third BP measurement (home)	17 yr +
HAZCK5MI	HAZA8CK5	K5 for third BP measurement (home)	17 yr +
HOUSEHOLD YOUTH QUESTIONNAIRE ITEMS			
HYD1MI	HYD1	How is health of SP in general	2 mo–16 yr
HYF2MI	HYF2	Condition of natural teeth	2 yr–16 yr
BONE DENSITOMETRY			
BDPFNDMI	BDPFNBMD	Bone mineral density of femur neck-gm/cm**2	20 yr +
BDPINDMI	BDPINBMD	BMD of intertrochanter region-gm/cm**2	20 yr +
BDPKMI	BDPK	K value for scan	20 yr +
BDPTOAMI	BDPTOARE	Bone area of total region - cm **2	20 yr +
BDPTODMI	BDPTOBMD	Bone mineral density total region-gm/cm**2	20 yr +
BDPTRDMI	BDPTRBMD	BMD of trochanter region - gm/cm**2	20 yr +
BDPWTDMI	BDPWTBMD	BMD of Ward's triangle region-gm/cm**2	20 yr +
BODY MEASUREMENTS			
BMPBUTMI	BMPBUTTO	Buttocks circumference (cm)	2 yr +
BMPHEAMI	BMPHEAD	Head circumference (cm)	2 mo–7 yr
BMPHTMI	BMPHT	Standing height (cm)	2 yr +
BMPKNEMI	BMPKNEE	Knee height (cm)	60 yr +
BMPRECMI	BMPRECUM	Recumbent length (cm)	2 mo–3 yr
BMPSTHMI	BMPSITHT	Sitting height (cm)	2 yr +
BMPSB1MI	BMPSUB1	First subscapular skinfold (mm)	2 mo +
BMPSB2MI	BMPSUB2	Second subscapular skinfold (mm)	2 mo +
BMPSP1MI	BMPSUP1	First suprailiac skinfold (mm)	2 yr +
BMPSP2MI	BMPSUP2	Second suprailiac skinfold (mm)	2 yr +
BMPTR1MI	BMPTRI1	First triceps skinfold (mm)	2 mo +
BMPTR2MI	BMPTRI2	Second triceps skinfold (mm)	2 mo +
BMPWSTMI	BMPWAIST	Waist circumference (cm)	2 yr +
BMPWTMI	BMPWT	Weight (kg)	2 mo +

Table 1 (continued): Variables selected for imputation in the NHANES III Multiply Imputed Data Set, and the non-imputed variables in previously released NHANES III public-use files to which they correspond

MI name	Previously	Description	Age range
FUNDUS PHOTOGRAPHY			
FPPSUDMI	FPPSUDRU	Summary drusen score	40 yr +
FPPSUMMI	FPPSUMAC	Summary age-related maculopathy score	40 yr +
FPPSURMI	FPPSURET	Summary diabetic retinopathy score	40 yr +
BLOOD AND URINE ASSAY ITEMS			
FEPMI	FEP	Serum iron (ug/dl)	1 yr +
FRPMI	FRP	Ferritin (ng/ml)	1 yr +
HDPMI	HDP	Serum HDL cholesterol (mg/dL)	4 yr +
HGPMI	HGP	Hemoglobin (g/dl)	1 yr +
HTPMI	HTP	Hematocrit (%)	1 yr +
MCPSIMI	MCPSI	Mean cell hemoglobin: SI	1 yr +
MHPMI	MHP	Mean cell hemoglobin concentration (g/dl)	1 yr +
MVPSIMI	MVPSI	Mean cell volume: SI (fl)	1 yr +
PBPMI	PBP	Lead (ug/dl)	1 yr +
PHPFSTMI	PHPFAST	Length of calculated fast (in hours)	1 yr +
PXPMI	PXP	Serum transferrin saturation (%)	1 yr +
RCPMI	RCP	Red blood cell count (x 10**6)	1 yr +
RWPMI	RWP	Red cell distribution width (%)	1 yr +
SEPMI	SEP	Selenium (ng/ml)	12 yr +
TCPMI	TCP	Serum cholesterol (mg/dL)	4 yr +
TGPMI	TGP	Serum triglycerides (mg/dL)	4 yr +
TIPMI	TIP	Serum TIBC (ug/dl)	1 yr +
REPLICATE BLOOD PRESSURE FROM MEC EXAMINATION			
PEP6G1MI	PEP6G1	K1, systolic, for 1st BP (mmHg)	5 yr +
PEP6G2MI	PEP6G2	K4, diastolic, for 1st BP(mmHg)	5 yr–19 yr
PEP6G3MI	PEP6G3	K5, diastolic, for 1st BP (mmHg)	5 yr +
PEP6H1MI	PEP6H1	K1, systolic, for 2nd BP (mmHg)	5 yr +
PEP6H2MI	PEP6H2	K4, diastolic, for 2nd BP(mmHg)	5 yr–19 yr
PEP6H3MI	PEP6H3	K5, diastolic, for 2nd BP (mmHg)	5 yr +
PEP6I1MI	PEP6I1	K1, systolic, for 3rd BP (mmHg)	5 yr +
PEP6I2MI	PEP6I2	K4, diastolic, for 3rd BP(mmHg)	5 yr–19 yr
PEP6I3MI	PEP6I3	K5, diastolic, for 3rd BP (mmHg)	5 yr +

Table 2: Age classes for imputation with number of interviewed and MEC-examined persons in each class

	Age class	Interviewed	Examined
1.	Newborn (under 1 year)	2,107	1,961
2.	1 year old	1,339	1,258
3.	2–3 years old	2,536	2,388
4.	4–7 years old	3,426	3,225
5.	8–16 years old	4,536	4,281
6.	17–19 years old	1,225	1,132
7.	20–39 years old	7,377	6,836
8.	40–59 years old	4,852	4,435
9.	60+ years old	6,596	5,302
	Total	33,994	30,818

2 Imputation models

2.1 Multivariate linear models with random effects

Within each age class, the model used to create imputations was a multivariate extension of a linear random-effects regression commonly applied to longitudinal and clustered data. Random-effects models describe responses that are intercorrelated because the units of observation are nested or grouped within larger units. In NHANES III, intercorrelations tend to arise because of the survey's multistage design. In particular, similarities may be expected among sampled persons from the same survey location. Accounting for intercorrelations within these locations is important because statistical methods recommended for the analysis of NHANES III—methods appropriate for data from complex surveys—rely heavily upon variation across these locations to calculate standard errors. For an imputation procedure to be compatible with these analysis procedures, appropriate levels of variation should be preserved both within and across locations.

Sampled persons in NHANES III came from 89 survey locations. Location indicators are regarded as confidential and are not released to the public, either in the NHANES III Multiply Imputed Data Set or in other public use data files. For this reason, the imputation procedures described here cannot be duplicated by researchers outside of the National Center for Health Statistics.

Before describing the multivariate linear random-effects model, we first review the univariate version. Let $y_i = (y_1, y_2, \ldots, y_{n_i})^T$ represent the vector of responses for a single variable for subjects $j = 1, 2, \ldots, n_i$ within cluster i, $i = 1, \ldots, N$. Suppose that these responses follow a linear regression of the form

$$y_i \;=\; X_i\beta + Z_i b_i + \varepsilon_i, \tag{1}$$

where X_i $(n_i \times p)$ and Z_i $(n_i \times q)$ are matrices of covariates, β contains regression coefficients common to all clusters, and b_i contains coefficients specific to cluster i. In popular terminology, β and b_i are called 'fixed effects' and 'random effects,' respectively. The random effects are assumed to be drawn from a multivariate normal population, $b_1, \ldots, b_N \sim N(0, \Psi)$, and the elements of ε_i are independent normal residuals, $\varepsilon_i \sim N(0, \sigma^2 I)$. Taken together, these distributional assumptions imply that

$$y_i \;\sim\; N(\, X_i\beta,\, Z_i\Psi Z_i^T + \sigma^2 I).$$

Models of this type were proposed by Hartley and Rao (1967) and popularized by Laird and Ware (1982), Jennrich and Schluchter (1986), Bryk and Raudenbush (1992) and others. Procedures for fitting these models are now found in major statistical packages including PROC MIXED (Littell et al., 1996) from SAS (SAS Institute Inc., 1999), S-PLUS (Mathsoft, Inc., 1997), and STATA (Stata Corporation, 1997). The columns of X_i usually include a constant term for an intercept and covariates describing the individuals and the cluster. The columns of Z_i, which are usually a subset of the columns of X_i, may include a constant term and subject-level covariates whose effects on the response may randomly vary by cluster. Setting $Z_i = (1, \ldots, 1)^T$ produces a random-intercepts model with an intracluster correlation of $\rho = \Phi/(\sigma^2 + \Phi)$.

The random-effects model described above could potentially be used to predict and impute a single variable in a cluster survey. But to jointly impute many variables at once and preserve correlations among them, the model must be extended to multivariate responses. Suppose that a set of variables $Y_1, Y_2, \ldots, Y_r$ is jointly measured for subjects $j = 1, \ldots, n_i$ in cluster i. The data for this cluster may be arranged as a matrix with one column for each variable

and one row for each subject,

$$y_i = \begin{bmatrix} y_{i11} & y_{i12} & \cdots & y_{i1r} \\ y_{i21} & y_{i22} & \cdots & y_{i2r} \\ \vdots & \vdots & \ddots & \vdots \\ y_{in_i1} & y_{in_i2} & \cdots & y_{in_ir} \end{bmatrix},$$

where y_{ijk} denotes the value of variable Y_k for subject j. The model for y_i is

$$y_i = X_i\beta + Z_ib_i + \varepsilon_i, \tag{2}$$

where X_i $(n_i \times p)$ and Z_i $(n_i \times q)$ contain covariates, β contains fixed effects and b_i contains random effects. Although (2) has the same appearance as (1), it is now a multivariate regression; β and b_i are now matrices with r columns, one column for predicting each of the variables $Y_1, Y_2, \ldots, Y_r$, and ε_i is a matrix with the same dimensions as y_i $(n_i \times r)$. The random effects and residuals are assumed to be distributed as

$$\text{vec}(b_i) \sim N(0, \Psi), \tag{3}$$

$$\text{vec}(\varepsilon_i) \sim N(0, (\Sigma \otimes I)), \tag{4}$$

where 'vec' denotes the vectorization of a matrix by stacking its columns. The covariance matrix Ψ in (3) has dimension $qr \times qr$, and the Kronecker product notation in (4) indicates that the rows of ε_i are independently distributed as $N(0, \Sigma)$, where Σ is $r \times r$. Note that in this multivariate model, all of the covariates in X_i and Z_i appear as predictors for each of the columns of y_i. The coefficients for the response variables contained in the r columns of β and b_i will vary, but the same set of predictors applies to each response. In this application to NHANES III, the matrices Z_i were set to $(1, \ldots, 1)^T$ for all clusters, producing random intercepts for each of the r response variables.

Multivariate random-effects regression models have received only limited attention in statistical literature. A model similar to (2) was considered by Reinsel (1984) who derived closed-form estimates with balanced data. Shah, Laird and Schoenfeld (1997) implemented an EM algorithm for unbalanced data in a bivariate ($r = 2$) setting with missing values in y_i. Schafer and Yucel (1999, under review) describe additional algorithms for parameter estimation and multiple imputation. These methods assume that the missing values are

missing at random in the sense described by Rubin (1976) and Little and Rubin (1987). The imputation procedures discussed by Schafer and Yucel (1999, under review) are implemented in a software library called PAN (Schafer, 1998) which operates in S-PLUS and can be downloaded from `http://www.stat.psu.edu/~jls/misoftwa.html`.

2.2 Response variables to be imputed

For each of the nine age groups shown in Table 2, a model of the form (2) was constructed for all interviewed persons in survey locations $i = 1, \ldots, 89$. Response variables in the columns of y_i included all of the variables in Table 1 applying to that age group, and in certain cases some additional variables potentially related to them. For example, the model for newborns (infants under one year) included response variables `DMPPIR`, `HFF1`, `HYD1`, `BMPHEAD`, `BMPRECUM`, `BMPSUB1`, `BMPSUB2`, `BMPTRI1`, `BMPTRI2`, and `BMPWT`. Across the nine models, the number of response variables ranged from $r = 10$ (newborns) to $r = 66$ (persons 60+).

Note that model (2) regards the r response variables as individually and jointly normally distributed within subgroups defined by the covariates in X_i and Z_i. A best, this assumption is only approximately satisfied. The distributions of many of the variables listed in Table 1 are substantially skewed. To produce imputations whose distributions resemble those of the observed data, many of the response variables were transformed by standard power functions such as the logarithm, square root, or reciprocal square root; modeling and imputation were carried out on the transformed data, and after imputation the variables were transformed back to their original scales. As a final step, the continuously distributed imputed values were rounded to the same precision found in the observed data. For example, blood pressure readings (mm Hg) are recorded in NHANES III as even integers, so imputed blood pressure readings were rounded to even integers.

In several instances, power transformations that approximately removed skewness did not produce satisfactory results. For example, some of the skin-fold measurements, after being transformed to near-symmetry, still exhibited lighter-than-normal tails; imputing these under a normal model might have produced unusually low or high imputations outside the realm of physical plausibility. These problematic variables were transformed by a method based

on the empirical cumulative distribution function (cdf) which forced them to approximate normality. Suppose $y_1, \ldots, y_n$ denotes a sample of numbers. Let $r_1, \ldots, r_n$ denote the integer ranks (lowest to highest) for these numbers, with tied values being assigned the average rank among the ties. Define the empirical cdf as $F(y_i) = r_i/(n+1)$. Finally, let Φ denote the standard normal cdf and Φ^{-1} the standard normal quantile function (e.g. $\Phi(1.96) = .95$ and $\Phi^{-1}(.95) = 1.96$). The transformed values

$$y_i^* \;=\; \Phi^{-1}(\, F(y_i)\,),;\; i = 1, \ldots, n \tag{5}$$

will tend to be approximately normally distributed regardless of the distribution of $y_1, \ldots, y_n$. We will refer to (5) as the 'empirical normal transformation.' When the empirical normal transformation was applied to a variable, the imputed values of y^* were transformed back by $y_i = F^{-1}(\,\Phi(y_i^*)\,)$. Imputing in this manner tends to preserve distributional shape quite well in an overall sense, but it produces duplication of extreme values rather than a smooth continuum in the tails.

Several of the variables listed in Table 1 are binary or ordinal scales (e.g. self-reported health status `HAB1`, which takes values from 1=excellent to 5=poor). These variables were included in the imputation models without transformation, and the imputed values were rounded to the nearest category. Normal based imputation and rounding of binary and ordinal variables has been shown to perform quite well in a variety of simulation studies (e.g. Schafer, 1997, chap. 6).

Regardless of the method used—a power transformation, the empirical normal method, or no transformation at all—imputed values in the NHANES III Multiply Imputed Data Set may not accurately reflect extreme tail behavior for many variables. For this reason, users are advised not to use these data for statistical analyses that are sensitive to extreme values, e.g. estimation of 98th percentiles. In fact, none of the NHANES III public release data sets may produce reliable inferences regarding extreme tail behavior; this is an inherent limitation of the NHANES III sample size, not the imputation method. For analyses about less extreme aspects of distributional shape—e.g. the estimation of means, medians, quartiles, or 10th and 90th percentiles–the imputation procedure is expected to perform well.

2.3 Model covariates

Many analyses of NHANES III variables are carried out within cross-classifications by age, sex, and race/ethnicity. To produce accurate national estimates within these subgroups, each of the nine imputation models incorporated this essential demographic information. Indicator variables for gender, race/ethnicity (coded as African-American, Mexican-American, and other), and a linear term for age were included in the columns of X_i, along with their two- and three-way products.

Another important covariate appearing in each model was the logarithm of household size. Household size, along with race/ethnicity and age, affected the probability that an individual was selected into the NHANES III sample. Household size is also strongly related to rates of nonresponse. Including this variable in the imputation models helps to eliminate systematic biases in the imputed values that could arise from over-sampling and differential response rates.

Finally, additional items from the household family, youth and adult questionnaires were used to define model covariates. These items, which are listed in Table 3, served as predictors in the imputation models but were not themselves imputed. These variables were chosen in consultation with subject matter experts at the National Center for Health Statistics either because (a) they might be related for obvious medical or physiological reasons to the response variables in Table 1, or because (b) they are likely to appear in variety of secondary analyses by users of NHANES III data. Examples of (a) include 'Have you ever been told that you have high blood pressure?' (`HAE2`) and 'Are you currently taking prescribed medication for high blood pressure?' (`HAE5A`), which may obviously be related to blood pressure readings. Examples of (b) include years of education (`HFA7`, `HFA8`) and marital status (`HFA12`). Some of these variables could could not be used for certain age groups because they indicate conditions that are extremely rare for the age group in question. For example, 'Have you ever been told that you have osteoporosis?' (`HAG11`) could only appear in the model for persons of age 60+ because virtually no positive responses to this question were seen in any other age group.

Some of the covariates listed in Table 3 had minor amounts of missing values. These

Table 3: Household interview variables in the NHANES III Multiply Imputed Data Set that served as potential predictors in the imputation models but were not imputed

Name	Description	Age range
HOUSEHOLD FAMILY QUESTIONNAIRE ITEMS		
HFA7	Highest grade or yr of school attended	2 mo +
HFA8	Finished highest grade/yr attended	2 mo +
HFA12	Marital status	14 yr +
HOUSEHOLD YOUTH QUESTIONNAIRE ITEMS		
HYE1G	Doc ever say had asthma	2 mo-16 yr
HYE1H	Doc ever say had chronic bronchitis	2 mo-16 yr
HYE6A	Doc ever say had high blood pressure	4 yr-16 yr
HYE6B	Doc ever say had high blood cholesterol	4 yr-16 yr
HYE15	Has ever had anemia	2 mo-16 yr
HYH2	Have trouble seeing w/one or both eyes	3 yr-16 yr
HYH10	Ever had troub hearing w/1 or both ears	2 mo-16 yr
HOUSEHOLD ADULT QUESTIONNAIRE ITEMS		
HAC1A	Ever told had arthritis	17 yr +
HAC1B	Which type of arthritis	17 yr +
HAC1C	Ever told had congestive heart failure	17 yr +
HAC1D	Ever told had stroke	17 yr +
HAC1E	Ever told had asthma	17 yr +
HAC1F	Ever told had chronic bronchitis	17 yr +
HAC1G	Ever told had emphysema	17 yr +
HAC1H	Ever told had hay fever	17 yr +
HAC1I	Ever told had cataracts	17 yr +
HAC1J	Ever told had goiter	17 yr +
HAC1K	Ever told had thyroid disease	17 yr +
HAC1L	Ever told had lupus	17 yr +
HAC1M	Ever told had gout	17 yr +
HAC1N	Ever told had skin cancer	17 yr +
HAC1O	Ever told had other type of cancer	17 yr +
HAD1	Ever told had diabetes	17 yr +
HAE2	Ever told had high blood pressure	17 yr +
HAE4A	Ever told to take prescr med for HBP	17 yr +
HAE4B	Ever told to ctrl/lose wt for HBP	17 yr +
HAE5A	Now taking prescr med for HBP	17 yr +
HAE5B	Is now ctrl/lose wt for HBP	17 yr +
HAE6	Ever had blood cholesterol checked	17 yr +
HAE7	Ever told had high cholesterol	17 yr +
HAF1	Ever had chest pain/discomfort	17 yr +
HAF10	Ever told had heart attack	17 yr +
HAG2	Ever had back pain most days for 1 mo	20 yr +
HAG3	Have back pain in past 12 months	20 yr +
HAG5A	Ever told had fractured hip	20 yr +
HAG5B	Ever told had fractured wrist	20 yr +
HAG5C	Ever told had fractured spine	20 yr +
HAG11	Ever told had osteoporosis	20 yr +

Table 3 (continued): Household interview variables in the NHANES III Multiply Imputed Data Set that served as potential predictors in the imputation models but were not imputed

Name	Description	Age range
HOUSEHOLD ADULT QUESTIONNAIRE ITEMS		
HAG12	Were treated for osteoporosis	20 yr +
HAN6HS	Beer and lite beer - times/month	17 yr +
HAN6IS	Wine, champagne - times/month	17 yr +
HAN6JS	Hard liquor - times/month	17 yr +
HAP1	Have total blindness	17 yr +
HAP1A	If yes, one or both eyes	17 yr +
HAP2	Use glasses, contacts, or both	17 yr +
HAP3	Trouble seeing with one or both eyes	17 yr +
HAP10	Have total deafness	17 yr +
HAP10A	If yes, one or both ears	17 yr +
HAR1	Smoked 100 cigarettes in life	17 yr +
HAR3	Smoke cigarettes now	17 yr +
HAR14	Used chewing tobacco, snuff	17 yr +
HAR16	Chew tobacco, snuff now	17 yr +
HAR23	Smoked 20 cigars in life	17 yr +
HAR24	Smoke cigars now	17 yr +
HAR26	Smoked 20 pipes of tobacco in life	17 yr +
HAR27	Smoke pipe now	17 yr +

missing values were handled in two ways. For any medical condition that was relatively rare in the NHANES III sample (e.g. thyroid disease), missing values were combined with negative responses into a single category. For other conditions that were not as rare, the variable was incorporated into the columns of y_i rather than X_i and treated as a response with missing values. In the latter situation, missing values for the variable were actually imputed, but because imputations were not generated consistently for all age classes, the imputed values were discarded.

2.4 Additional notes on model specification

As described above, the covariates in X_i for each model included a constant term for the intercept, columns producing main effects and interactions for gender $\times$ race/ethnicity $\times$ age, the logarithm of household size, and additional variables from Table 3. The number of columns in X_i varied from $p = 7$ (one year olds) to $p = 35$ (age 60+). With r response variables, the fixed effects in β form a $p \times r$ matrix. In the largest of the models (age 60+),

the number of regression coefficients being simultaneously estimated was $35 \times 66 = 2310$.

For each model, the matrix Z_i was simply a constant $(1, \ldots, 1)^T$ which allowed the intercepts to randomly vary by cluster. Under this specification, the random effects b_i were vectors of length r, with the jth element of b_i representing the deviation of the Y_j-intercept in survey location i from the population average Y_j-intercept. Allowing the intercepts to vary by survey location is consistent with earlier models which allowed a separate intercept for each location (Schafer, Khare, and Ezzati-Rice, 1993). With up to $r = 66$ response variables and only 89 locations, it was not possible to obtain stable estimates of covariances among all the elements of $b_i = (b_{i1}, \ldots, b_{ir})$. For this reason, the between-location covariance matrix Ψ was assumed to be a diagonal matrix, with the off-diagonal elements set to zero.

Note that the NHANES III sampling weights, which are determined by individuals' probabilities of being selected into the sample, played no formal role in fitting the models or imputing missing observations. However, all of the important determinants of selection probability (age, race/ethnicity, household size and survey location) were conditioned upon in all models, greatly reducing any possibility that the oversampling of certain groups in NHANES III could bias the imputations toward the characteristics of the overrepresented groups. Empirical evidence supporting this type of unweighted imputation modeling for an unequally weighted sample is provided by the simulation results of Little et al. (1995).

A complete listing of the response variables, transformation methods, and covariates appearing in each of the nine imputation models is provided in Appendix A, Tables A1–A9.

2.5 The missing-at-random assumption

Procedures used to create the NHANES III Multiply Imputed Data Set assume that the missing values are 'missing at random' (MAR) in the sense defined by Little and Rubin (1987) and Rubin (1987). Under MAR, the probability that any data value is missing may depend on quantities that are observed but not on quantities that are missing. Nearly all missing-data procedures applied to sample surveys assume some form of MAR or make assumptions that are even stronger. It is important to note that MAR is not an inherent property of any

data set; rather, it is a property of the data and the model used to describe them. The MAR assumption becomes more plausible as the model is enriched to include more information related to the nonresponse mechanism. In designing the imputation models for NHANES III, every attempt was made to incorporate variables related to response rates. Once these variables have been included, it is no longer possible to verify or refute the MAR assumption by examining rates and patterns of missing values (unless additional unverifiable assumptions are made). Further discussion of MAR and its practical implications is given by Schafer (1997, ch. 2).

3 Computational procedures

3.1 Gibbs sampler

The computational algorithm used to create multiple imputations is a Markov chain Monte Carlo (MCMC) procedure called a Gibbs sampler. MCMC is a class of simulation techniques especially useful in Bayesian statistical analyses. Various types of MCMC methods are reviewed in the volume edited by Gilks, Richardson & Spiegelhalter (1996). A gentle introduction to Gibbs sampling is provided by Casella and George (1992). The application of MCMC to multiple imputation is discussed by Schafer (1997).

Methods of Gibbs sampling for linear random-effects models have previously been published by Gelfand *et al.* (1990), Zeger and Karim (1991), and Carlin (1996). Those articles pertain to models for a single response variable. Schafer and Yucel (1999, under review) have extended the method to multiple response variables with incomplete data. This particular Gibbs sampler is based on the observation that the multivariate linear random-effects model has the following unknown components: the missing values in $y_1, y_2, \ldots, y_N$, the random effects $b_1, b_2, \ldots, b_N$, the fixed effects β, and the covariance matrices Σ and Ψ. For the purpose of imputation, we are interested only in simulating the missing data in $y_1, y_2, \ldots, y_N$; the other unknown quantities are merely a nuisance. To simulate the missing data properly, however, we must take into account the uncertainty in these other quantities and how it contributes to missing-data uncertainty. Expressing this uncertainty through mathematical

formulas is difficult, so we account for the interdependence among the unknown quantities through a process of iterative simulation.

The unknown quantities are simulated in a three-step cycle: (a) Random values of $b_1, b_2, \ldots, b_N$ are drawn based on some plausible assumed values for the missing data and the parameters β, Σ, and Ψ. (b) New random values of the unknown parameters β, Σ, and Ψ ae drawn based on the assumed values for the missing data and the values of $b_1, b_2, \ldots, b_N$ obtained in (a). (c) New random values for the missing data are drawn given the values of $b_1, b_2, \ldots, b_N$ obtained in (a) and the parameters obtained in (b). At the end of this cycle, the parameters and missing data from (b) and (c) become the values assumed in step (a) at the start of the next cycle. Repeating (a), (b), and (c) in turn defines a Markov chain, a sequence in which the distribution of the unknown quantities at any cycle depends on their simulated values at the previous cycle. The state of the process at cycle 2 may be strongly correlated with its state at cycle 1, but at subsequent cycles $3, 4, 5, \ldots$ the relationship to the original state weakens. When a sufficient number of cycles have been taken to make the resulting state essentially independent of the original state, then the process is said to have 'converged' or 'achieved stationarity.' Upon convergence, the final simulated values for the missing data have in fact come from the distribution from which multiple imputations should be drawn. Specific formulas for steps (a)–(c) are given by Schafer and Yucel (1999, under review).

This algorithm may used to create M simulated versions of the complete data in the following way. Starting with some plausible initial values, run the Gibbs sampler for k cycles where k is large enough to ensure convergence, and take the final simulated version of the missing data as the first imputation; then return to the original set of starting values, run the Gibbs sampler (using a new random-number generator seed) for another k cycles, and take the final simulated version of the missing data as the second imputation; and so on. This method requires M runs of length k cycles each. Another and perhaps more convenient way is to perform one long run of Mk cycles, saving the simulated values of the missing data after cycle $k, 2k, \ldots, Mk$ as the M imputations. The latter method differs from the former only in that the final values from each subchain of length k become the starting values for the next

subchain of length k.

3.2 Convergence issues

Convergence of an MCMC procedure means convergence to a probability distribution rather than convergence to a set of fixed values. To say that the algorithm has converged by k cycles actually means that the random state of the process at cycle $t + k$ is statistically independent of its state at cycle t for $t = 1, 2, \ldots$. After running the Gibbs sampler, one may examine the output stream over many cycles to see how many are needed to achieve this independence. Suppose that we collect and store the simulated values for one parameter θ (a particular element of β, Ψ, or Σ) over a large number C of consecutive cycles. These values $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(C)}$ can be regarded as a time series. The lag-k autocorrelation, which is the correlation between pairs $\theta^{(t)}$ and $\theta^{(t+k)}$ $(t = 1, 2, \ldots, C - k)$, can be calculated for various values of k to determine how large k must be for the correlations to die down. In principle, one should examine autocorrelations for each parameter in the model and identify a value of k large enough to guarantee that the lag-k autocorrelations for all parameters are effectively zero. Experience with real data indicates that the greatest levels of serial dependence are almost always seen in variance and covariance parameters, and in particular within the elements of Ψ. It is usually sufficient to monitor the behavior of the elements of Ψ because it is with respect to these parameters that the algorithm tends to converge the most slowly. For more discussion on monitoring the convergence of MCMC algorithms, see Schafer (1997, chap. 4).

The speed at which the Gibbs sampler converges in this application is influenced by a combination of factors pertaining to the data and the model. First, it is affected by the amounts and patterns of missing data in the matrices $y_1, y_2, \ldots, y_N$; high rates of missing information lead to slower convergence. It is also affected by our ability to estimate the cluster-level random effects $b_1, b_2, \ldots, b_N$; if estimates of these are highly variable, then convergence is slowed. Finally, convergence behavior is influenced by the number of clusters N. As the number of clusters grows, the distribution of the random Ψ matrix at each cycle becomes more tightly concentrated around the sample covariance matrix of $b_1, b_2, \ldots, b_N$ from

the previous cycle. As this distribution becomes tighter, the elements of Ψ are less free to wander away from their values at the previous cycle, producing higher correlations from one cycle to the next and thus slowing convergence.

Fortunately, in this particular application, all of the factors mentioned above tend to favor rapid convergence. The per-variable missingness rates in NHANES III are moderately low (less than 30% among interviewed persons). The number of clusters ($N = 89$) is not large enough to severely restrict the variability of individual elements of Ψ at each iteration. Within each cluster, the sample size is sufficiently large to obtain accurate estimates of the cluster-specific means, producing stable estimates for random effects $b_1, \ldots, b_N$. Time-series and autocorrelation plots from preliminary runs of the Gibbs sampler revealed no serial dependence in parameters beyond lag 30 for any age group. Therefore, it appeared that $k = 30$ cycles between imputations was sufficient to produce imputations that were essentially independent. For an extra margin of safety, $k = 50$ cycles between imputations were taken for the larger age groups, and in the smaller age groups where cycles could be executed very quickly, the number was increased to $k = 100$.

3.3 Prior distributions

To apply this Gibbs sampler, one must specify Bayesian prior distributions for the unknown model parameters β, Ψ and Σ. Bayesian procedures treat unknown parameters as random variables and assign prior probability distributions to them to reflect one's knowledge or belief about the parameters before the data are seen. A modern overview of Bayesian modeling and computation is provided by Gelman et al. (1995). Some statisticians tend to prefer Bayesian procedures on principle, whereas others avoid them on principle. We hold a pragmatic view, accepting the prior distribution as a mathematical device which allows us to generate the imputations in a principled fashion. In many applications, it is desirable to use prior distributions that are weak or highly dispersed, reflecting a state of relative ignorance about model parameters. Weak priors tend to minimize the subjective influence of the prior, allowing the observed data to speak for themselves.

Following common practice, we assume a noninformative, improper uniform prior distri-

bution for β over the real space $\mathcal{R}^{rp}$; for the covariance matrices Ψ and Σ, however, proper prior distributions must be applied to guarantee existence of the joint posterior distribution (Hobart and Casella, 1996). The prior distribution most commonly applied to a covariance matrix is the inverted Wishart distribution. With an inverted Wishart prior, the user must provide (a) an a priori estimate or guess for the matrix in question, and (b) a number for the degrees of freedom on which this prior estimate or guess is based. To specify the prior distributions, we first calculated the variance among the observed values for each response variable in the model. Prior guesses for Ψ and Σ were then derived by supposing that both matrices were diagonal and that the overall variance for each variable was split equally among the within-cluster and between-cluster components. The prior degrees of freedom were set to the minimum numbers required to ensure that the prior distribution is proper, making the prior as 'weak' as possible.

3.4 Futher computational details

The Gibbs sampling procedures described above were carried out using the PAN library (Schafer, 1998) within the statistical package S-PLUS (Mathsoft, 1997). For efficiency, the computationally intensive operations in PAN are implemented in Fortran. Imputations were created on a single 400 Mhz Pentium II personal computer with 128 MB of memory in less than three hours per age class.

4 Graphical comparisons of observed and imputed values

4.1 Marginal comparisons

One way to evaluate the quality of an imputation procedure is to compare the distributions of imputed and nonimputed values for each variable to see if they are similar. Such comparisons should be interpreted with caution. Discrepancies between the distributions of imputed and nonimputed values do not necessarily reveal a shortcoming of the imputation procedure, because individuals with missing values may systematically differ from those with observed values in a variety of ways. For example, suppose that the probability of nonresponse is

higher for elderly persons than for the non-elderly; if an imputation procedure is working properly, then the imputed values should more closely resemble those of elderly persons than the overall sample. Many types of systematic differences between observed and imputed values are allowed under the missing-at-random (MAR) assumption.

Nevertheless, graphical comparisons between observed and imputed values are useful for detecting gross problems. Imputed values should not lie outside the range of physical plausibility. If the systematic differences between respondents and nonrespondents are not unusually strong, then the distributions of observed and imputed values should be similar in location, scale and shape. Comparisons of the marginal distributions of observed and imputed values are provided in Appendix B. For each variable, side-by-side histograms display the observed values and the imputed values from imputation sets 1, 2, and 3 (results for sets 4 and 5 are similar to those from 1–3 and are not shown). For the most part, important distributional features are preserved remarkably well. Patterns of skewness and even bimodality in the observed data are usually evident in the imputed values. For example, many of the body measurement variables that were collected for both adults and children reveal two distinct modes; in each case the imputed values show the same bimodal pattern. The combination of age-specific models and nonlinear transformations appears to be quite effective for preserving important aspects of distributional shape.

4.2 Bivariate comparisons

In addition to providing quality imputations for each variable, the NHANES III multiple imputation procedures were also designed to preserve important relationships among variables. A representative selection of bivariate scatterplots of observed and imputed values is provided in Appendix C. For each pair of variables in question, three scatterplots are shown. The first plot displays all individuals for which both variables were observed. The second plot displays individuals for which one or both variables were imputed, using the first set of imputed values. The third plot displays the same set of individuals using the second set of imputed values. Results for imputation sets 3–5 are similar to those from 1–2 and are not shown. Examination of these scatterplots suggest that the imputation procedures do preserve

essential features of inter-variable relationships.

Users of the NHANES III Multiply Imputed Data Set are encouraged to explore the data and produce additional graphical displays comparing observed and imputed data relevant to their analyses. Imputation flags provided in the data set allow the user to easily distinguish imputed values from observed ones. Details on file formats, imputed variables, and imputation flags are provided in the documentation files accompanying the NHANES III Multiply Imputed Data Set.

References

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models.* Newbury Park, CA: Sage.

Carlin, B.P. (1996) Hierarchical longitudinal modelling. In *Markov Chain Monte Carlo in Practice,* W.R. Gilks, S. Richardson and D.J. Spiegelhalter (eds.), London, U.K.: Chapman & Hall, pp. 303–319.

Casella, G. and George, E.I. (1992) Explaining the Gibbs sampler. *The American Statistician*, 46, 167–174.

Department of Health and Human Services (DHHS) (1994) *Plan and Operation of the Third National Health and Nutrition Examination Survey*, 1988–94. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Department of Health and Human Services (DHHS) (1996) *NHANES III Reference Manuals and Reports.* CD-ROM. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Department of Health and Human Services (DHHS) (1997) *National Health and Nutrition Examination Survey, III*, 1988–1994. CD-ROM, Series 11, No. 1A, ASCII Version. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Department of Health and Human Services (DHHS) (1998) *National Health and Nutrition Examination Survey, III*, 1988–1994. CD-ROM, Series 11, No. 2A, ASCII Version. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Gelfand, A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M. (1990) "Illustration of Bayesian inference in normal data models using Gibbs sampling," *Journal of the American Statistical Association,* 85, 972–985.

Gelman, G., Carlin J.B., Stern, H.S., and Rubin, D.B. (1995) *Bayesian Data Analysis,* London: Chapman & Hall.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (Eds.). (1996) *Markov Chain Monte Carlo in Practice*. Lndon: Chapman & Hall.

Hartley, H.O. and Rao, J.N.K. (1967) Maximum likelihood estimation for the mixed analysis of variance model. *Biometrics*, 54, 93–108.

Hobart, J.P. and Casella, G. (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461–1473.

Jennrich, R.I. and Schluchter, M.D. (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 38, 967–974.

Laird, N.M. and Ware, J.H. (1982) Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996) *SAS System for Mixed Models,* Cary, NC: SAS Institute, Inc.

Little, R.J.A. (1986) Survey nonresponse adjustments for estimation of means. *International Statistical Review*, 54, 139–157.

Little, R.J.A., Ezzati-Rice, T.M., Johnson, W., Khare, M., Rubin, D.B. and Schafer, J.L. (1995) A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. *Proceedings of the Annual Research Conference*, 257–266. Washington, DC: Department of Commerce, Bureau of the Census. Included with the NHANES III Multiple Imputation Research Data Set (DHHS, 2000, CD-ROM).

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data,* New York: Wiley.

MathSoft, Inc. (1997) *S-PLUS User's guide,* Data Analysis Product Division, Seattle, WA: MathSoft, Inc.

Reinsel G. (1984) Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association,* 77, 190–195

Rubin, D.B. (1976) Inference and missing data. *Biometrika,* 63, 581–592.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys,* New York: Wiley.

SAS Institute, Inc. (1999) *SAS/STAT User's Guide, Version 8,* Cary, NC: SAS Institute, Inc.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. (1998) *PAN: Multiple imputation for longitudinal and clustered data under a multivariate linear mixed model*, software library for S-PLUS. Written in S-PLUS and Fortran-77. Available at `http://www.stat.psu.edu/~jls/`.

Schafer J.L., Khare, M. and Ezzati-Rice, T.M. (1993) Multiple imputation of missing data in NHANES III. Proceedings of the Annual Research Conference, 459-487. Washington, DC: Department of Commerce, Bureau of the Census. Included with the NHANES III Multiple Imputation Research Data Set (DHHS, 2000, CD-ROM).

Schafer, J.L. and Yucel, R.M. (under review) Multivariate linear mixed-effects models with missing values. Submitted to *Journal of Computational and Graphical Statistics.*

Shah, A., Laird, N., Schoenfeld, D. (1997) A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association,* 92, 775–779

Stata Corporation (1997) *Stata Reference Manual,* College Station, TX: Stata Press.

Zeger, S.L. and Karim, M.R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79–86.

Appendix A: Details of imputation models

Table A1 (a): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 1: NEWBORNS (UNDER ONE YEAR)

Variable	Transformation	Post-imputation processing
BMPHEAD	none	round to nearest 0.1
BMPRECUM	none	round to nearest 0.1
BMPSUB1	log	antilog, round to nearest 0.1
BMPSUB2	log	antilog, round to nearest 0.1
BMPTRI1	log	antilog, round to nearest 0.1
BMPTRI2	log	antilog, round to nearest 0.1
BMPWT	log	antilog, round to nearest 0.01
DMPPIR	$y^{1/2}$	square, round to nearest 0.001
HFF1	none	round to 1 or 2
HYD1	none	round to 1,2,3,4,5

Table A1 (b): Covariates appearing in NHANES III imputation model with fixed effects

CLASS 1: NEWBORNS (UNDER ONE YEAR)

Covariate	Description
`constant`	one
`hhs.log`	log of household size
`age`	age in months
`sex`	indicator for male
`race1`	indicator for Black
`race2`	indicator for Mexican-American
`age.sex`	product of `age`, `sex`
`age.race1`	product of `age`, `race1`
`age.race2`	product of `age`, `race2`
`sex.race1`	product of `sex`, `race1`
`sex.race2`	product of `sex`, `race2`
`age.sex.race1`	product of `age`, `sex`, `race1`
`age.sex.race2`	product of `age`, `sex`, `race2`

Table A2 (a): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 2: ONE YEAR OLDS

Variable	Transformation	Post-imputation processing
BMPHEAD	none	round to nearest 0.1
BMPRECUM	none	round to nearest 0.1
BMPSUB1	log	antilog, round to nearest 0.1
BMPSUB2	log	antilog, round to nearest 0.1
BMPTRI1	log	antilog, round to nearest 0.1
BMPTRI2	log	antilog, round to nearest 0.1
BMPWT	log	antilog, round to nearest 0.01
DMPPIR	$y^{1/2}$	square, round to nearest 0.001
FEP	$y^{1/2}$	square, round to nearest integer
FRP	log	antilog, round to nearest integer
HFF1	none	round to 1 or 2
HGP	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
HTP	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
HYD1	none	round to 1,2,3,4,5
HYE15	none	imputed values discarded
MCPSI	none	round to nearest 0.01
MHP	none	round to nearest 0.01
MVPSI	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
PBP	log	antilog, round to nearest 0.01
PHPFAST	$y^{1/2}$	square, round to nearest 0.01
PXP	$y^{1/2}$	square, round to nearest 0.1
RCP	y^2	$\max(y,0)^{1/2}$, round to nearest 0.01
RWP	y^{-2}	$\max(y,0)^{-1/2}$, round to nearest 0.01
TIP	none	round to nearest integer

Table A2 (b): Covariates appearing in NHANES III imputation model with fixed effects

CLASS 2: ONE YEAR OLDS

Covariate	Description
`constant`	one
`hhs.log`	log of household size
`sex`	indicator for male
`race1`	indicator for Black
`race2`	indicator for Mexican-American
`sex.race1`	product of `sex`, `race1`
`sex.race2`	product of `sex`, `race2`

Table A3 (a): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 3: 2–3 YEARS OLD

Variable	Transformation	Post-imputation processing
BMPBUTTO	log	antilog, round to nearest 0.1
BMPHEAD	none	round to nearest 0.1
BMPHT	none	round to nearest 0.1
BMPRECUM	none	round to nearest 0.1
BMPSITHT	none	round to nearest 0.1
BMPSUB1	log	antilog, round to nearest 0.1
BMPSUB2	log	antilog, round to nearest 0.1
BMPSUP1	log	antilog, round to nearest 0.1
BMPSUP2	log	antilog, round to nearest 0.1
BMPTRI1	log	antilog, round to nearest 0.1
BMPTRI2	log	antilog, round to nearest 0.1
BMPWAIST	log	antilog, round to nearest 0.1
BMPWT	log	antilog, round to nearest 0.01
DMPPIR	$y^{1/2}$	square, round to nearest 0.001
FEP	$y^{1/2}$	square, round to nearest integer
FRP	log	antilog, round to nearest integer
HFF1	none	round to 1 or 2
HGP	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
HTP	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
HYD1	none	round to 1,2,3,4,5
HYE15	none	imputed values discarded
HYF2	none	round to 1,2,3,4,5,6
MCPSI	none	round to nearest 0.01
MHP	none	round to nearest 0.01
MVPSI	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
PBP	log	antilog, round to nearest 0.01
PHPFAST	$y^{1/2}$	square, round to nearest 0.01
PXP	$y^{1/2}$	square, round to nearest 0.1
RCP	y^2	$\max(y,0)^{1/2}$, round to nearest 0.01
RWP	y^{-2}	$\max(y,0)^{-1/2}$, round to nearest 0.01
TIP	none	round to nearest integer

Table A3 (b): Covariates appearing in NHANES III imputation model with fixed effects

CLASS 3: 2–3 YEARS OLD

Covariate	Description
`constant`	one
`hhs.log`	log of household size
`age`	age in years
`sex`	indicator for male
`race1`	indicator for Black
`race2`	indicator for Mexican-American
`age.sex`	product of `age`, `sex`
`age.race1`	product of `age`, `race1`
`age.race2`	product of `age`, `race2`
`sex.race1`	product of `sex`, `race1`
`sex.race2`	product of `sex`, `race2`
`age.sex.race1`	product of `age`, `sex`, `race1`
`age.sex.race2`	product of `age`, `sex`, `race2`

Table A4 (a): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 4: 4–7 YEARS OLD

Variable	Transformation	Post-imputation processing
BMPBUTTO	log	antilog, round to nearest 0.1
BMPHEAD	none	round to nearest 0.1
BMPHT	none	round to nearest 0.1
BMPSITHT	none	round to nearest 0.1
BMPSUB1	empirical normal	inverse empirical normal
BMPSUB2	empirical normal	inverse empirical normal
BMPSUP1	empirical normal	inverse empirical normal
BMPSUP2	empirical normal	inverse empirical normal
BMPTRI1	empirical normal	inverse empirical normal
BMPTRI2	empirical normal	inverse empirical normal
BMPWAIST	log	antilog, round to nearest 0.1
BMPWT	log	antilog, round to nearest 0.01
DMPPIR	$y^{1/2}$	square, round to nearest 0.001
FEP	$y^{1/2}$	square, round to nearest integer
FRP	log	antilog, round to nearest integer
HDP	none	round to nearest integer
HFF1	none	round to 1 or 2
HGP	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
HTP	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
HYD1	none	round to 1,2,3,4,5
HYE15	none	imputed values discarded
HYE1G	none	imputed values discarded
HYF2	none	round to 1,2,3,4,5,6
MCPSI	none	round to nearest 0.01
MHP	none	round to nearest 0.01
MVPSI	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
PBP	log	antilog, round to nearest 0.01
PEP6G1	none	round to even integer, set to missing for age 4
PEP6G2	none	round to even integer, missing for age 4
PEP6G3	none	round to even integer, missing for age 4

Table A4 (a) (continued): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 4: 4–7 YEARS OLD

Variable	Transformation	Post-imputation processing
PEP6H1	none	round to even integer, set to missing for age 4
PEP6H2	none	round to even integer, missing for age 4
PEP6H3	none	round to even integer, missing for age 4
PEP6I1	none	round to even integer, set to missing for age 4
PEP6I2	none	round to even integer, missing for age 4
PEP6I3	none	round to even integer, missing for age 4
PHPFAST	$y^{1/2}$	square, round to nearest 0.01
PXP	$y^{1/2}$	square, round to nearest 0.1
RCP	y^2	$\max(y,0)^{1/2}$, round to nearest 0.01
RWP	y^{-2}	$\max(y,0)^{-1/2}$, round to nearest 0.01
TCP	log	antilog, round to nearest integer
TGP	log	antilog, round to nearest integer
TIP	none	round to nearest integer

Table A4 (b): Covariates appearing in NHANES III imputation model with fixed effects

CLASS 4: 4–7 YEARS OLD

Covariate	Description
`constant`	one
`hhs.log`	log of household size
`age`	age in years
`sex`	indicator for male
`race1`	indicator for Black
`race2`	indicator for Mexican-American
`age.sex`	product of `age`, `sex`
`age.race1`	product of `age`, `race1`
`age.race2`	product of `age`, `race2`
`sex.race1`	product of `sex`, `race1`
`sex.race2`	product of `sex`, `race2`
`age.sex.race1`	product of `age`, `sex`, `race1`
`age.sex.race2`	product of `age`, `sex`, `race2`

Table A5 (a): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 5: 8–16 YEARS OLD

Variable	Transformation	Post-imputation processing
BMPBUTTO	log	antilog, round to nearest 0.1
BMPHT	none	round to nearest 0.1
BMPSITHT	none	round to nearest 0.1
BMPSUB1	empirical normal	inverse empirical normal
BMPSUB2	empirical normal	inverse empirical normal
BMPSUP1	empirical normal	inverse empirical normal
BMPSUP2	empirical normal	inverse empirical normal
BMPTRI1	empirical normal	inverse empirical normal
BMPTRI2	empirical normal	inverse empirical normal
BMPWAIST	log	antilog, round to nearest 0.1
BMPWT	log	antilog, round to nearest 0.01
DMPPIR	$y^{1/2}$	square, round to nearest 0.001
FEP	$y^{1/2}$	square, round to nearest integer
FRP	log	antilog, round to nearest integer
HDP	log	antilog, round to nearest integer
HFF1	none	round to 1 or 2
HGP	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
HTP	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
HYD1	none	round to 1,2,3,4,5
HYE15	none	imputed values discarded
HYE1G	none	imputed values discarded
HYF2	none	round to 1,2,3,4,5,6
MCPSI	none	round to nearest 0.01
MHP	none	round to nearest 0.01
MVPSI	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
PBP	log	antilog, round to nearest 0.01
PEP6G1	none	round to even integer
PEP6G2	none	round to even integer
PEP6G3	none	round to even integer
PEP6H1	none	round to even integer

Table A5 (a) (continued): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 5: 8–16 YEARS OLD

Variable	Transformation	Post-imputation processing
PEP6H2	none	round to even integer
PEP6H3	none	round to even integer
PEP6I1	none	round to even integer
PEP6I2	none	round to even integer
PEP6I3	none	round to even integer
PHPFAST	$y^{1/2}$	square, round to nearest 0.01
PXP	$y^{1/2}$	square, round to nearest 0.1
RCP	y^2	$\max(y,0)^{1/2}$, round to nearest 0.01
RWP	y^{-2}	$\max(y,0)^{-1/2}$, round to nearest 0.01
SEP	log	antilog, round to integer, make missing for age 8–11
TCP	log	antilog, round to nearest integer
TGP	log	antilog, round to nearest integer
TIP	none	round to nearest integer

Table A5 (b): Covariates appearing in NHANES III imputation model with fixed effects

CLASS 5: 8–16 YEARS OLD

Covariate	Description
`constant`	one
`hhs.log`	log of household size
`age`	age in years
`sex`	indicator for male
`race1`	indicator for Black
`race2`	indicator for Mexican-American
`age.sex`	product of `age`, `sex`
`age.race1`	product of `age`, `race1`
`age.race2`	product of `age`, `race2`
`sex.race1`	product of `sex`, `race1`
`sex.race2`	product of `sex`, `race2`
`age.sex.race1`	product of `age`, `sex`, `race1`
`age.sex.race2`	product of `age`, `sex`, `race2`

Table A6 (a): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 6: 17–19 YEARS OLD

Variable	Transformation	Post-imputation processing
BMPBUTTO	y^{-3}	$\max(y,0)^{-1/3}$, round to nearest 0.1
BMPHT	none	round to nearest 0.1
BMPSITHT	none	round to nearest 0.1
BMPSUB1	empirical normal	inverse empirical normal
BMPSUB2	empirical normal	inverse empirical normal
BMPSUP1	empirical normal	inverse empirical normal
BMPSUP2	empirical normal	inverse empirical normal
BMPTRI1	empirical normal	inverse empirical normal
BMPTRI2	empirical normal	inverse empirical normal
BMPWAIST	y^{-1}	y^{-1}, round to nearest 0.1
BMPWT	$y^{-1/2}$	y^{-2}, round to nearest 0.01
COLLEGE[a]	none	imputed values discarded
DMPPIR	$y^{1/2}$	square, round to nearest 0.001
FEP	$y^{1/2}$	square, round to nearest integer
FRP	log	antilog, round to nearest integer
HAB1	none	round to 1,2,3,4,5
HAM5	none	round to nearest integer
HAM6	$y^{-1/2}$	y^{-2}, round to nearest integer
BRWNLQ[b]	none	round to 0,1,2
HAQ1	none	round to 1,2,3,4,5,6
CIGARETT[c]	none	round to 0 or 1
HAT28	reorder as 1,4,2,3	round to nearest integer, restore original order
HAZA8AK1	none	round to even integer
HAZA8AK5	none	round to even integer
HAZA8BK1	none	round to even integer
HAZA8BK5	none	round to even integer
HAZA8CK1	none	round to even integer
HAZA8CK5	none	round to even integer
HDP	$y^{1/2}$	square, round to nearest integer
HFF1	none	round to 1 or 2
HGP	none	round to nearest 0.01

[a] Defined as 1 if HFA7 $\geq$ 13, 2 if HFA7 $\leq$ 12.

[b] Defined as 0 if han6hs + han6is + han6js = 0, 1 if 1 $\leq$ han6hs + han6is + han6js $\leq$ 10, 2 if han6hs + han6is + han6js $>$ 10.

[c] Defined as 0 if HAR1 = 2, 1 if HAR3 = 1 and HAR1 $\neq$ 2.

Table A6 (a) (continued): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 6: 17–19 YEARS OLD

Variable	Transformation	Post-imputation processing
HTP	none	round to nearest 0.01
MARRIED[d]	none	imputed values discarded
MCPSI	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
MHP	y^2	$\max(y,0)^{1/2}$, round to nearest 0.01
MVPSI	y^3	$\max(y,0)^{1/3}$, round to nearest 0.01
PBP	log	antilog, round to nearest 0.01
PEP6G1	none	round to even integer
PEP6G2	none	round to even integer
PEP6G3	none	round to even integer
PEP6H1	none	round to even integer
PEP6H2	none	round to even integer
PEP6H3	none	round to even integer
PEP6I1	none	round to even integer
PEP6I2	none	round to even integer
PEP6I3	none	round to even integer
PHPFAST	$y^{1/2}$	square, round to nearest 0.01
PXP	$y^{1/2}$	square, round to nearest 0.1
RCP	none	round to nearest 0.01
RWP	y^{-5}	$\max(y,0)^{-1/5}$, round to nearest 0.01
SEP	$y^{1/2}$	square, round to nearest integer
TCP	log	antilog, round to nearest integer
TGP	log	antilog, round to nearest integer
TIP	$y^{1/2}$	square, round to nearest integer

[d] Defined as 1 if HFA12 ≤ 3, 2 otherwise.

Table A6 (b): Covariates appearing in NHANES III imputation model with fixed effects

CLASS 6: 17–19 YEARS OLD

Covariate	Description
constant	one
hhs.log	log of household size
age	age in years
sex	indicator for male
race1	indicator for Black
race2	indicator for Mexican-American
age.sex	product of age, sex
age.race1	product of age, race1
age.race2	product of age, race2
sex.race1	product of sex, race1
sex.race2	product of sex, race2
age.sex.race1	product of age, sex, race1
age.sex.race2	product of age, sex, race2
asthma	1 if HAC1E=2, 0 if HAC1E=1 or missing
hayfev	1 if HAC1H=2, 0 if HAC1H=1 or missing
sayhibp	1 if HAE2=2, 0 if HAE2=1 or missing
chestpn	1 if HAF1=2, 0 if HAF1=1 or missing

Table A7 (a): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 7: 20–39 YEARS OLD

Variable	Transformation	Post-imputation processing
BDPFNBMD	log	antilog, round to nearest 0.001
BDPINBMD	log	antilog, round to nearest 0.001
BDPK	empirical normal	inverse empirical normal
BDPTOARE	log	antilog, round to nearest 0.01
BDPTOBMD	log	antilog, round to nearest 0.001
BDPTRBMD	log	antilog, round to nearest 0.001
BDPWTBMD	$y^{1/2}$	square, round to nearest 0.001
BMPBUTTO	y^{-1}	y^{-1}, round to nearest 0.1
BMPHT	none	round to nearest 0.1
BMPSITHT	none	round to nearest 0.1
BMPSUB1	empirical normal	inverse empirical normal
BMPSUB2	empirical normal	inverse empirical normal
BMPSUP1	empirical normal	inverse empirical normal
BMPSUP2	empirical normal	inverse empirical normal
BMPTRI1	empirical normal	inverse empirical normal
BMPTRI2	empirical normal	inverse empirical normal
BMPWAIST	$y^{-1/2}$	y^{-2}, round to nearest 0.1
BMPWT	$y^{-1/2}$	y^{-2}, round to nearest 0.01
COLLEGE[a]	none	imputations discarded
DMPPIR	$y^{1/2}$	square, round to nearest 0.001
FEP	none	round to nearest integer
FRP	empirical normal	inverse empirical normal
HAB1	none	round to 1,2,3,4,5
HAF1	none	round to 1 or 2
HAG2	none	round to 1 or 2
HAM5	none	round to nearest integer
HAM6	$y^{-1/2}$	y^{-2}, round to nearest integer
BRWNLQ[b]	none	round to 0,1,2
HAQ1	none	round to 1,2,3,4,5,6
CIGARETT[c]	none	round to 0 or 1
HAT28	reorder as 1,4,2,3	round to nearest integer, restore original order

[a] Defined as 1 if HFA7 $\geq$ 13, 2 if HFA7 $\leq$ 12.

[b] Defined as 0 if han6hs + han6is + han6js = 0, 1 if 1 $\leq$ han6hs + han6is + han6js $\leq$ 10, 2 if han6hs + han6is + han6js $>$ 10.

[c] Defined as 0 if HAR1 = 2, 1 if HAR3 = 1 and HAR1 $\neq$ 2.

Table A7 (a) (continued): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 7: 20–39 YEARS OLD

Variable	Transformation	Post-imputation processing
HAZA8AK1	y^{-1}	y^{-1}, round to even integer
HAZA8AK5	none	round to even integer
HAZA8BK1	y^{-1}	y^{-1}, round to even integer
HAZA8BK5	none	round to even integer
HAZA8CK1	y^{-1}	y^{-1}, round to even integer
HAZA8CK5	none	round to even integer
HDP	$y^{1/3}$	cube, round to nearest integer
HFF1	none	round to 1 or 2
HGP	y^2	$y^{1/2}$, round to nearest 0.01
HTP	y^2	$y^{1/2}$, round to nearest 0.01
MARRIED[d]	none	imputed values discarded
MCPSI	y^2	$\max(y,0)^{1/2}$, round to nearest 0.01
MHP	none	round to nearest 0.01
MVPSI	none	round to nearest 0.01
PBP	empirical normal	inverse empirical normal
PEP6G1	y^{-1}	y^{-1}, round to even integer
PEP6G3	none	round to even integer
PEP6H1	y^{-1}	y^{-1}, round to even integer
PEP6H3	none	round to even integer
PEP6I1	y^{-1}	y^{-1}, round to even integer
PEP6I3	none	round to even integer
PHPFAST	$y^{1/2}$	square, round to nearest 0.01
PXP	$y^{1/2}$	square, round to nearest 0.1
RCP	none	round to nearest 0.01
RWP	y^{-4}	$\max(y,0)^{-1/4}$, round to nearest 0.01
SEP	log	antilog, round to nearest integer
TCP	log	antilog, round to nearest integer
TGP	log	antilog, round to nearest integer
TIP	$y^{1/2}$	square, round to nearest integer

[d] Defined as 1 if HFA12 ≤ 3, 2 otherwise.

Table A7 (b): Covariates appearing in NHANES III imputation model with fixed effects

CLASS 7: 20–39 YEARS OLD

Covariate	Description
constant	one
hhs.log	log of household size
age	age in years
sex	indicator for male
race1	indicator for Black
race2	indicator for Mexican-American
age.sex	product of age, sex
age.race1	product of age, race1
age.race2	product of age, race2
sex.race1	product of sex, race1
sex.race2	product of sex, race2
age.sex.race1	product of age, sex, race1
age.sex.race2	product of age, sex, race2
arthritis	1 if HAC1A=2, 0 if HAC1A=1 or missing
asthma	1 if HAC1E=2, 0 if HAC1E=1 or missing
bronchitis	1 if HAC1F=2, 0 if HAC1F=1 or missing
hayfev	1 if HAC1H=2, 0 if HAC1H=1 or missing
diabetes	1 if HAD1=2, 0 if HAD1=1 or missing
sayhibp	1 if HAE2=2, 0 if HAE2=1 or missing
sayhichol	1 if HAE7=2, 0 if HAE7=1 or missing
fractr	1 if hag5a + hag5b + hag5c<6, 0 otherwise or if any of them are missing

Table A8 (a): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 8: 40–59 YEARS OLD

Variable	Transformation	Post-imputation processing
BDPFNBMD	log	antilog, round to nearest 0.001
BDPINBMD	log	antilog, round to nearest 0.001
BDPK	empirical normal	inverse empirical normal
BDPTOARE	$y^{1/2}$	square, round to nearest 0.01
BDPTOBMD	log	antilog, round to nearest 0.001
BDPTRBMD	log	antilog, round to nearest 0.001
BDPWTBMD	log	antilog, round to nearest 0.001
BMPBUTTO	y^{-1}	y^{-1}, round to nearest 0.1
BMPHT	none	round to nearest 0.1
BMPSITHT	none	round to nearest 0.1
BMPSUB1	empirical normal	inverse empirical normal
BMPSUB2	empirical normal	inverse empirical normal
BMPSUP1	empirical normal	inverse empirical normal
BMPSUP2	empirical normal	inverse empirical normal
BMPTRI1	empirical normal	inverse empirical normal
BMPTRI2	empirical normal	inverse empirical normal
BMPWAIST	log	antilog, round to nearest 0.1
BMPWT	log	antilog, round to nearest 0.01
COLLEGE[a]	none	imputations discarded
DMPPIR	$y^{1/2}$	square, round to nearest 0.001
FEP	$y^{1/2}$	square, round to nearest integer
FPPSUDRU	none	round to 0,1,2
FPPSUMAC	none	round to 0,1,2
FPPSURET	none	round to 0,1,2,3
FRP	empirical normal	inverse empirical normal
HAB1	none	round to 1,2,3,4,5
HAF1	none	round to 1 or 2
HAG2	none	round to 1 or 2
HAM5	none	round to nearest integer
HAM6	log	antilog, round to nearest integer
BRWNLQ[b]	none	round to 0,1,2

[a] Defined as 1 if HFA7 $\geq$ 13, 2 if HFA7 $\leq$ 12.

[b] Defined as 0 if han6hs + han6is + han6js = 0, 1 if 1 $\leq$ han6hs + han6is + han6js $\leq$ 10, 2 if han6hs + han6is + han6js $>$ 10.

Table A8 (a) (continued): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 8: 40–59 YEARS OLD

Variable	Transformation	Post-imputation processing
HAP2	none	round to 1,2,3,4
HAP3	none	round to 1 or 2
HAQ1	none	round to 1,2,3,4,5,6
CIGARETT[c]	none	round to 0 or 1
HAT28	reorder as 1,4,2,3	round to nearest integer, restore original order
HAZA8AK1	log	antilog, round to even integer
HAZA8AK5	none	round to even integer
HAZA8BK1	log	antilog, round to even integer
HAZA8BK5	none	round to even integer
HAZA8CK1	log	antilog, round to even integer
HAZA8CK5	none	round to even integer
HDP	log	antilog, round to nearest integer
HFF1	none	round to 1 or 2
HGP	y^2	$y^{1/2}$, round to nearest 0.01
HTP	y^2	$y^{1/2}$, round to nearest 0.01
MARRIED[d]	none	imputed values discarded
MCPSI	y^2	$\max(y,0)^{1/2}$, round to nearest 0.01
MHP	none	round to nearest 0.01
MVPSI	none	round to nearest 0.01
PBP	empirical normal	inverse empirical normal
PEP6G1	log	antilog, round to even integer
PEP6G3	none	round to even integer, set negatives to zero
PEP6H1	log	antilog, round to even integer
PEP6H3	none	round to even integer, set negatives to zero
PEP6I1	log	antilog, round to even integer
PEP6I3	none	round to even integer, set negatives to zero
PHPFAST	$y^{1/2}$	square, round to nearest 0.01
PXP	$y^{1/2}$	square, round to nearest 0.1
RCP	none	round to nearest 0.01
RWP	y^{-3}	$\max(y,0)^{-1/3}$, round to nearest 0.01
SEP	log	antilog, round to nearest integer

[c] Defined as 0 if HAR1 = 2, 1 if HAR3 = 1 and HAR1 $\neq$ 2.

[d] Defined as 1 if HFA12 $\leq$ 3, 2 otherwise.

Table A8 (a) (continued): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 8: 40–59 YEARS OLD

Variable	Transformation	Post-imputation processing
TCP	log	antilog, round to nearest integer
TGP	log	antilog, round to nearest integer
TIP	$y^{1/2}$	square, round to nearest integer

Table A8 (b): Covariates appearing in NHANES III imputation model with fixed effects

CLASS 8: 40–59 YEARS OLD

Covariate	Description
constant	one
hhs.log	log of household size
age	age in years
sex	indicator for male
race1	indicator for Black
race2	indicator for Mexican-American
age.sex	product of age, sex
age.race1	product of age, race1
age.race2	product of age, race2
sex.race1	product of sex, race1
sex.race2	product of sex, race2
age.sex.race1	product of age, sex, race1
age.sex.race2	product of age, sex, race2
arthritis	1 if HAC1A=2, 0 if HAC1A=1 or missing
asthma	1 if HAC1E=2, 0 if HAC1E=1 or missing
bronchitis	1 if HAC1F=2, 0 if HAC1F=1 or missing
hayfev	1 if HAC1H=2, 0 if HAC1H=1 or missing
diabetes	1 if HAD1=2, 0 if HAD1=1 or missing
sayhibp	1 if HAE2=2, 0 if HAE2=1 or missing
hbpmed	1 if HAE4A=2, 0 if HAE4A=1 or missing
losewt	1 if HAE4B=2, 0 if HAE4B=1 or missing
colchk	1 if HAE6=2, 0 if HAE6=1 or missing
sayhichol	1 if HAE7=2, 0 if HAE7=1 or missing
fractr	1 if hag5a + hag5b + hag5c<6, 0 otherwise or if any of them are missing

Table A9 (a): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 9: 60+ YEARS OLD

Variable	Transformation	Post-imputation processing
BDPFNBMD	$y^{1/2}$	square, round to nearest 0.001
BDPINBMD	$y^{1/2}$	square, round to nearest 0.001
BDPK	empirical normal	inverse empirical normal
BDPTOARE	none	round to nearest 0.01
BDPTOBMD	none	round to nearest 0.001
BDPTRBMD	none	round to nearest 0.001
BDPWTBMD	$y^{1/2}$	square, round to nearest 0.001
BMPBUTTO	log	antilog, round to nearest 0.1
BMPHT	none	round to nearest 0.1
BMPKNEE	empirical normal	inverse empirical normal
BMPSITHT	none	round to nearest 0.1
BMPSUB1	empirical normal	inverse empirical normal
BMPSUB2	empirical normal	inverse empirical normal
BMPSUP1	empirical normal	inverse empirical normal
BMPSUP2	empirical normal	inverse empirical normal
BMPTRI1	empirical normal	inverse empirical normal
BMPTRI2	empirical normal	inverse empirical normal
BMPWAIST	log	antilog, round to nearest 0.1
BMPWT	log	antilog, round to nearest 0.01
DMPPIR	$y^{1/2}$	square, round to nearest 0.001
FEP	$y^{1/2}$	square, round to nearest integer
FPPSUDRU	none	round to 0,1,2
FPPSUMAC	none	round to 0,1,2
FPPSURET	none	round to 0,1,2,3
FRP	empirical normal	inverse empirical normal
HAB1	none	round to 1,2,3,4,5
HAC1A	none	imputed values discarded
HAC1I	none	imputed values discarded
HAF1	none	round to 1 or 2
HAF10	none	imputed values discarded

Table A9 (a) (continued): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 9: 60+ YEARS OLD

Variable	Transformation	Post-imputation processing
HAG2	none	round to 1 or 2
HAM5	none	round to nearest integer
HAM6	log	antilog, round to nearest integer
BRWNLQ[a]	none	round to 0,1,2
HAQ1	none	round to 1,2,3,4,5,6
CIGARETT[b]	none	round to 0 or 1
HAT28	reorder as 1,4,2,3	round to nearest integer, restore original order
HAZA8AK1	log	antilog, round to even integer
HAZA8AK5	empirical normal	inverse empirical normal
HAZA8BK1	log	antilog, round to even integer
HAZA8BK5	empirical normal	inverse empirical normal
HAZA8CK1	log	antilog, round to even integer
HAZA8CK5	empirical normal	inverse empirical normal
HDP	log	antilog, round to nearest integer
HFF1	none	round to 1 or 2
HGP	$y^{3/2}$	$y^{2/3}$, round to nearest 0.01
HTP	y^2	$y^{1/2}$, round to nearest 0.01
MARRIED[c]	none	imputed values discarded
MCPSI	empirical normal	inverse empirical normal
MHP	none	round to nearest 0.01
MVPSI	none	round to nearest 0.01
PBP	empirical normal	inverse empirical normal
PEP6G1	log	antilog, round to even integer
PEP6G3	empirical normal	inverse empirical normal
PEP6H1	log	antilog, round to even integer
PEP6H3	empirical normal	inverse empirical normal
PEP6I1	log	antilog, round to even integer
PEP6I3	empirical normal	inverse empirical normal

[a] Defined as 0 if han6hs + han6is + han6js = 0, 1 if 1 $\leq$ han6hs + han6is + han6js $\leq$ 10, 2 if han6hs + han6is + han6js $>$ 10.

[b] Defined as 0 if HAR1 = 2, 1 if HAR3 = 1 and HAR1 $\neq$ 2.

[c] Defined as 1 if HFA12 $\leq$ 3, 2 otherwise.

Table A9 (a) (continued): Response variables in NHANES III imputation model, transformations applied prior to imputation, and post-imputation processing of imputed values

CLASS 9: 60+ YEARS OLD

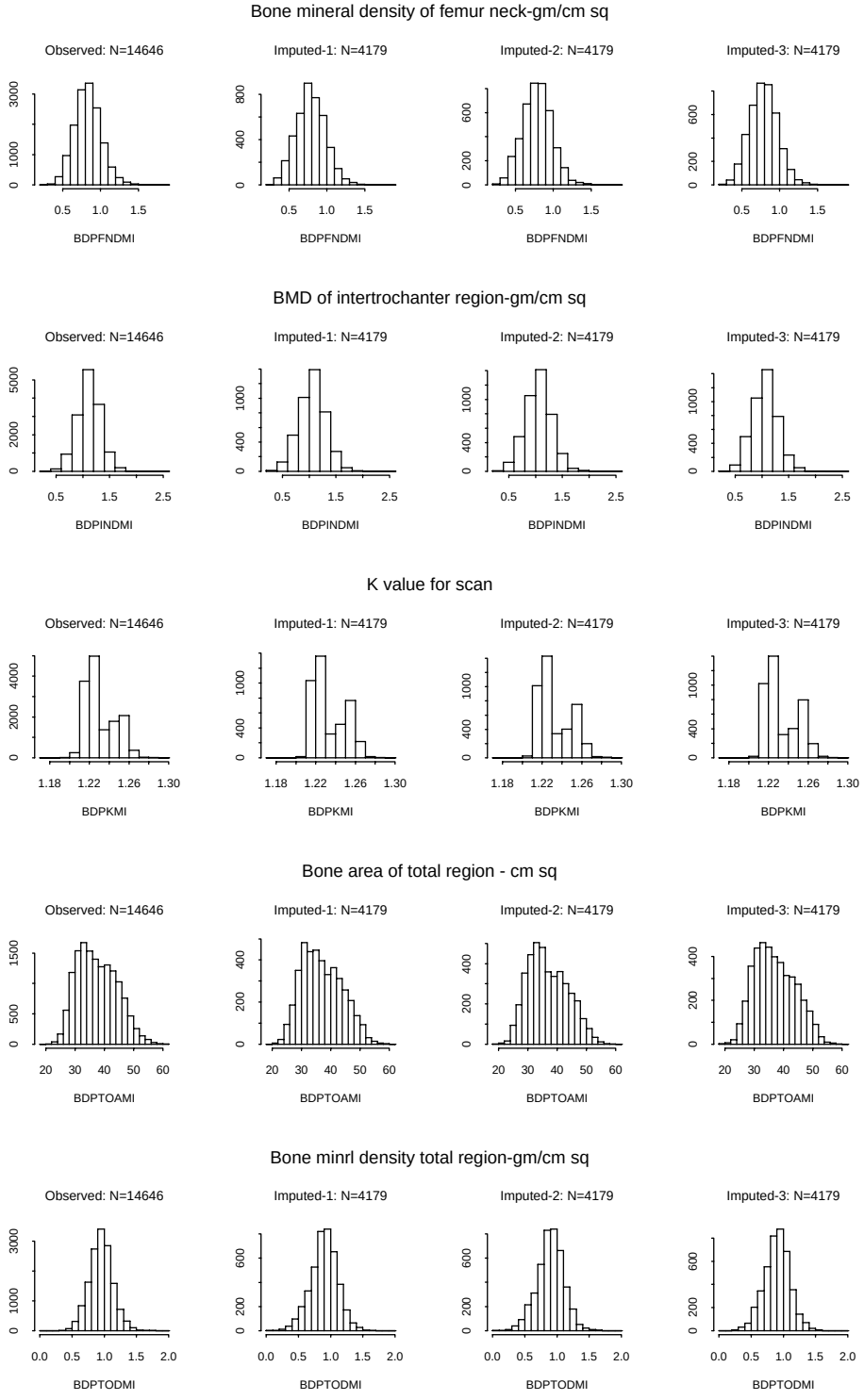
Variable	Transformation	Post-imputation processing
PHPFAST	$y^{1/2}$	square, round to nearest 0.01
PXP	$y^{1/2}$	square, round to nearest 0.1
RCP	none	round to nearest 0.01
RWP	y^{-3}	$\max(y,0)^{-1/3}$, round to nearest 0.01
SEP	log	antilog, round to nearest integer
TCP	$y^{1/2}$	square, round to nearest integer
TGP	log	antilog, round to nearest integer
TIP	none	round to nearest integer

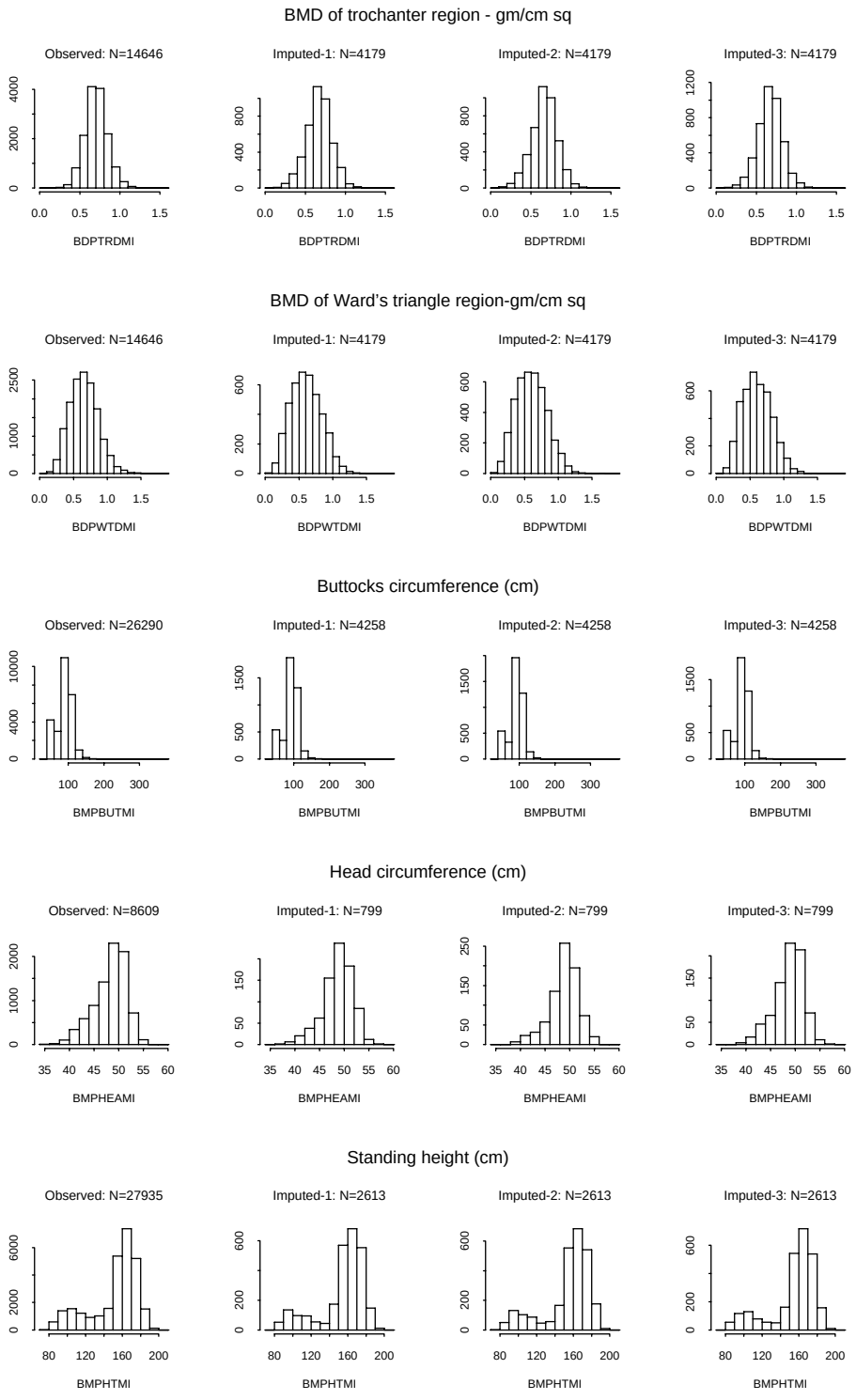
Table A9 (b): Covariates appearing in NHANES III imputation model with fixed effects

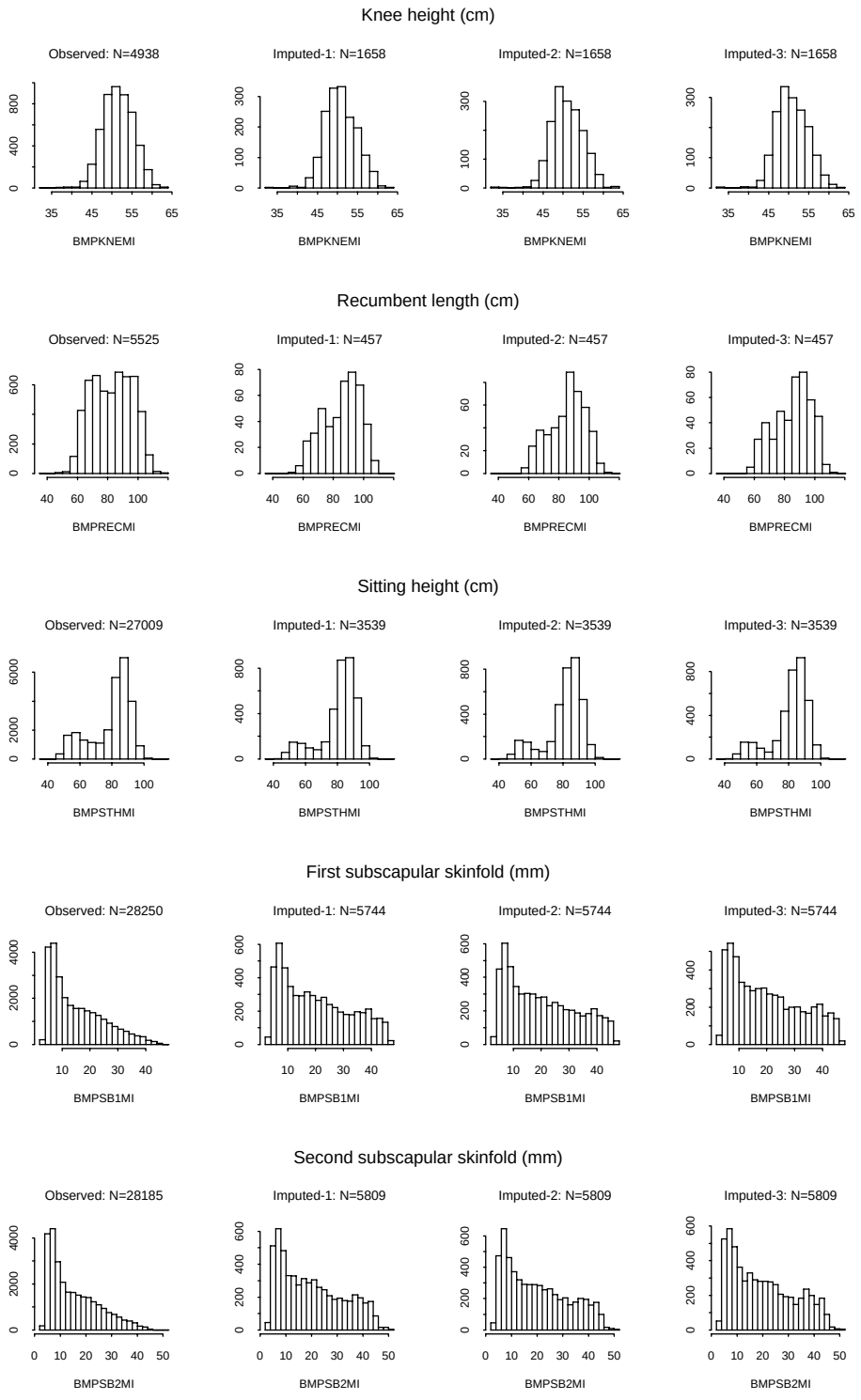
CLASS 9: 60+ YEARS OLD

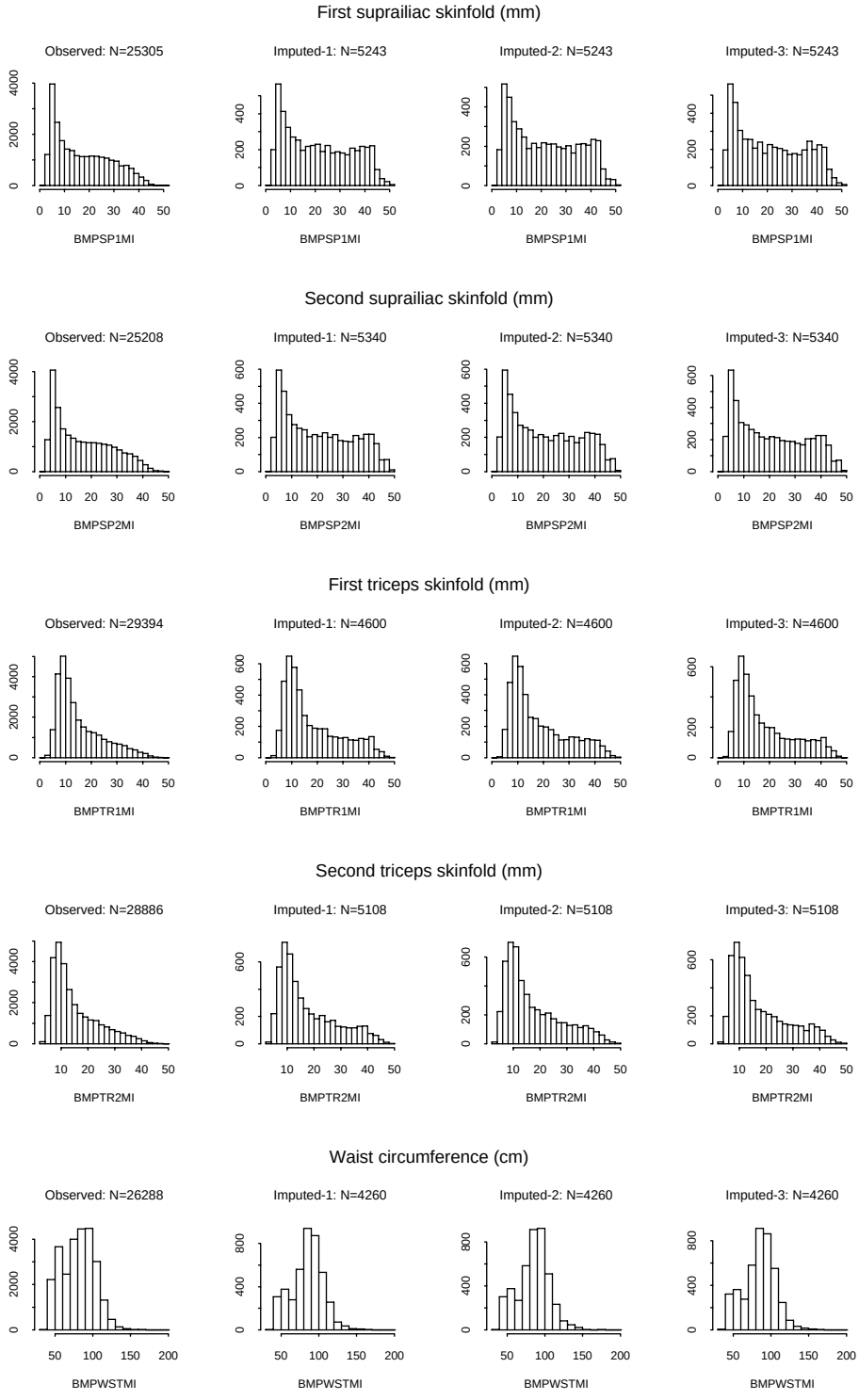
Covariate	Description
constant	one
hhs.log	log of household size
age	age in years
sex	indicator for male
race1	indicator for Black
race2	indicator for Mexican-American
age.sex	product of age, sex
age.race1	product of age, race1
age.race2	product of age, race2
sex.race1	product of sex, race1
sex.race2	product of sex, race2
age.sex.race1	product of age, sex, race1
age.sex.race2	product of age, sex, race2
hrtfail	1 if HAC1C=2, 0 if HAC1C=1 or missing
stroke	1 if HAC1D=2, 0 if HAC1D=1 or missing
asthma	1 if HAC1E=2, 0 if HAC1E=1 or missing
bronchitis	1 if HAC1F=2, 0 if HAC1F=1 or missing
emphysema	1 if HAC1G=2, 0 if HAC1G=1 or missing
hayfev	1 if HAC1H=2, 0 if HAC1H=1 or missing
goiter	1 if HAC1J=2, 0 if HAC1J=1 or missing
thyroid	1 if HAC1K=2, 0 if HAC1K=1 or missing
gout	1 if HAC1M=2, 0 if HAC1M=1 or missing
skincancer	1 if HAC1N=2, 0 if HAC1N=1 or missing
cancer	1 if HAC1O=2, 0 if HAC1O=1 or missing
diabetes	1 if HAD1=2, 0 if HAD1=1 or missing
hibpyes	1 if HAE2=1, 0 if HAE2=2 or missing
hibpno	1 if HAE2=2, 0 if HAE2=1 or missing
hbpmed	1 if HAE4A=2, 0 if HAE4A=1 or missing
losewt	1 if HAE4B=2, 0 if HAE4B=1 or missing
colchk	1 if HAE6=2, 0 if HAE6=1 or missing
sayhichol	1 if HAE7=2, 0 if HAE7=1 or missing
fractr	1 if hag5a + hag5b + hag5c<6, 0 otherwise or if any of them are missing
osteoporosis	1 if HAG11=2, 0 if HAG11=1 or missing
glasses	1 if HAP2≤3, 0 if HAP2=4 or missing
troubleseeing	1 if HAP3=1 and not missing, 0 otherwise

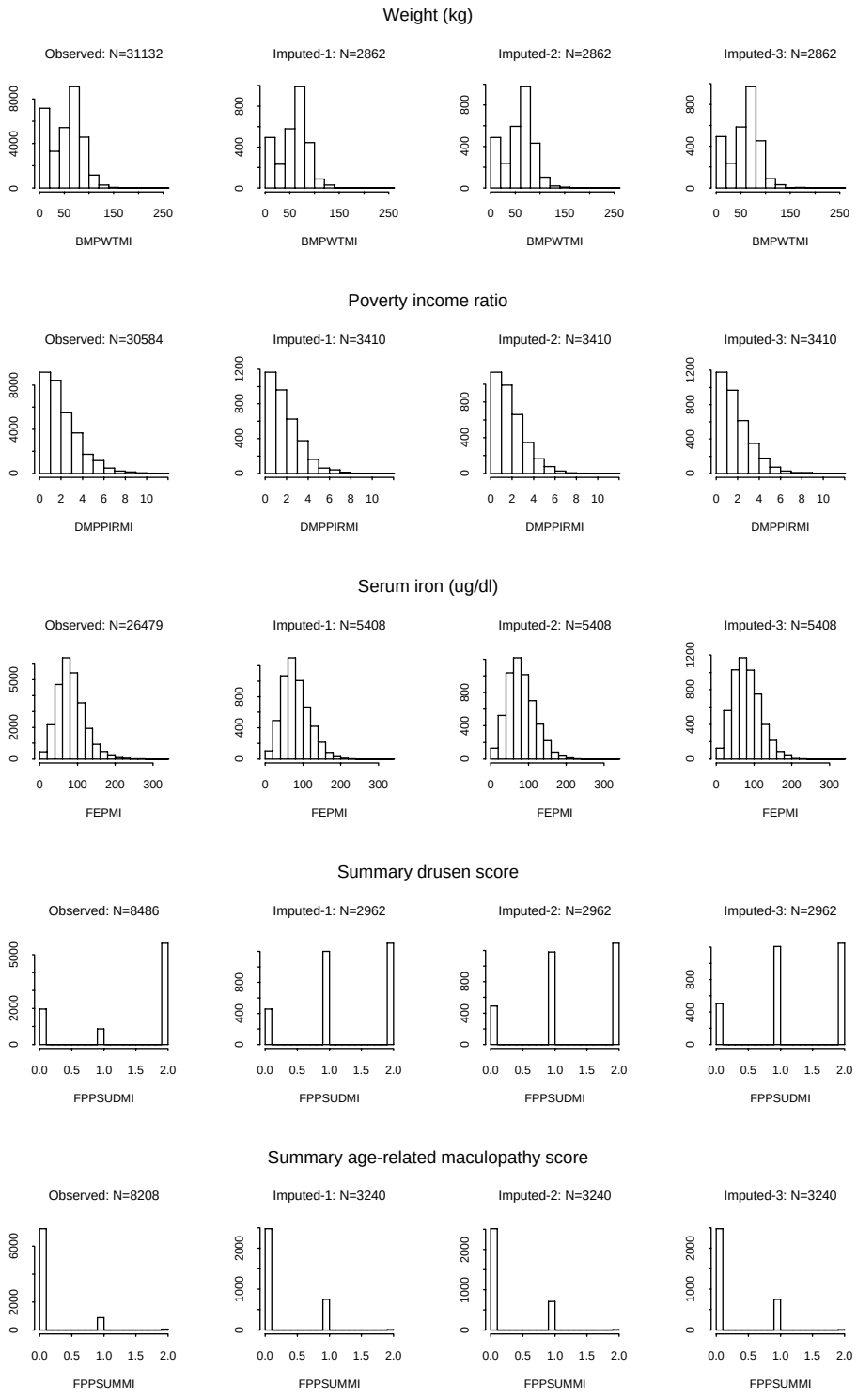
Appendix B: Comparisons of marginal distributions

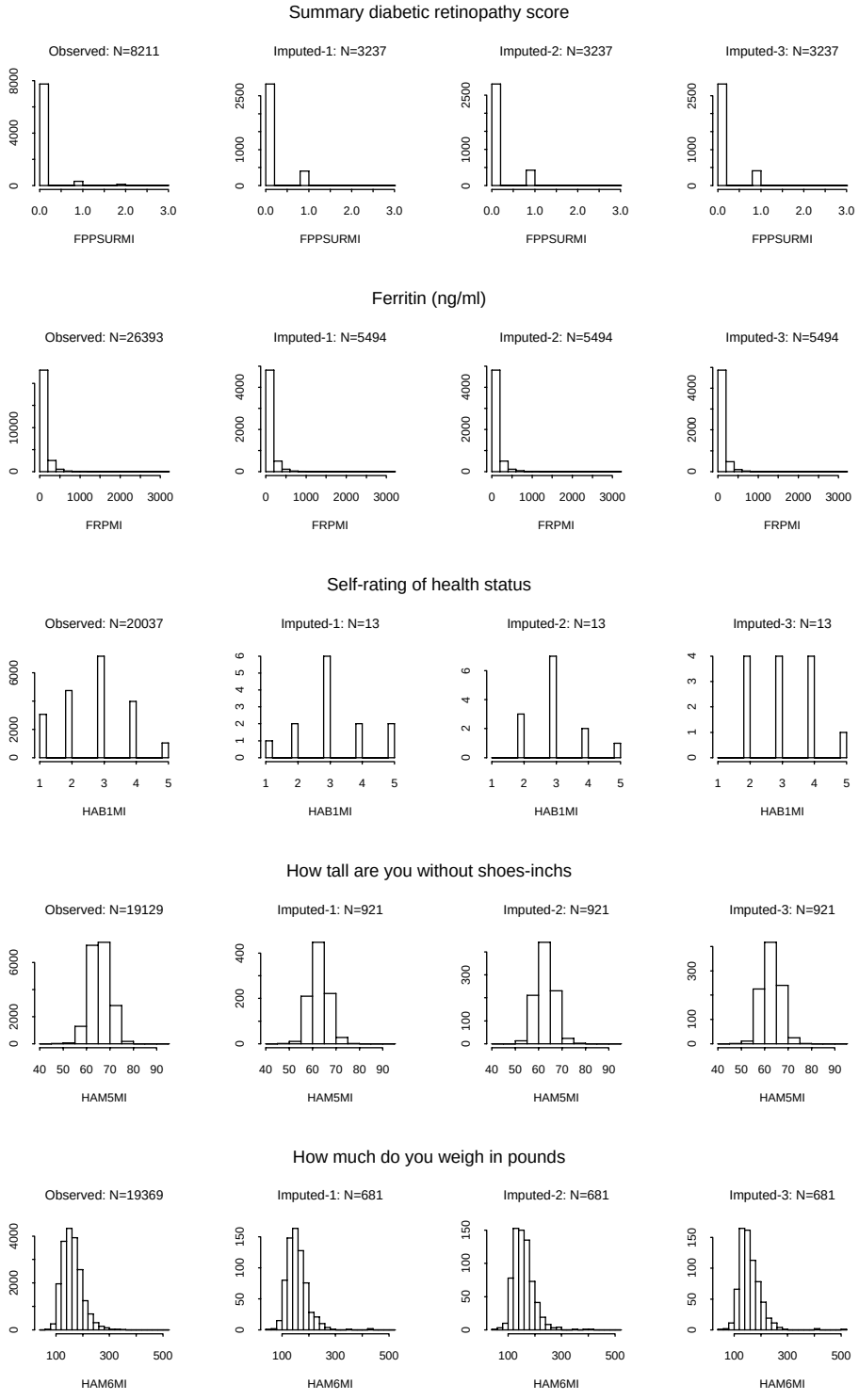


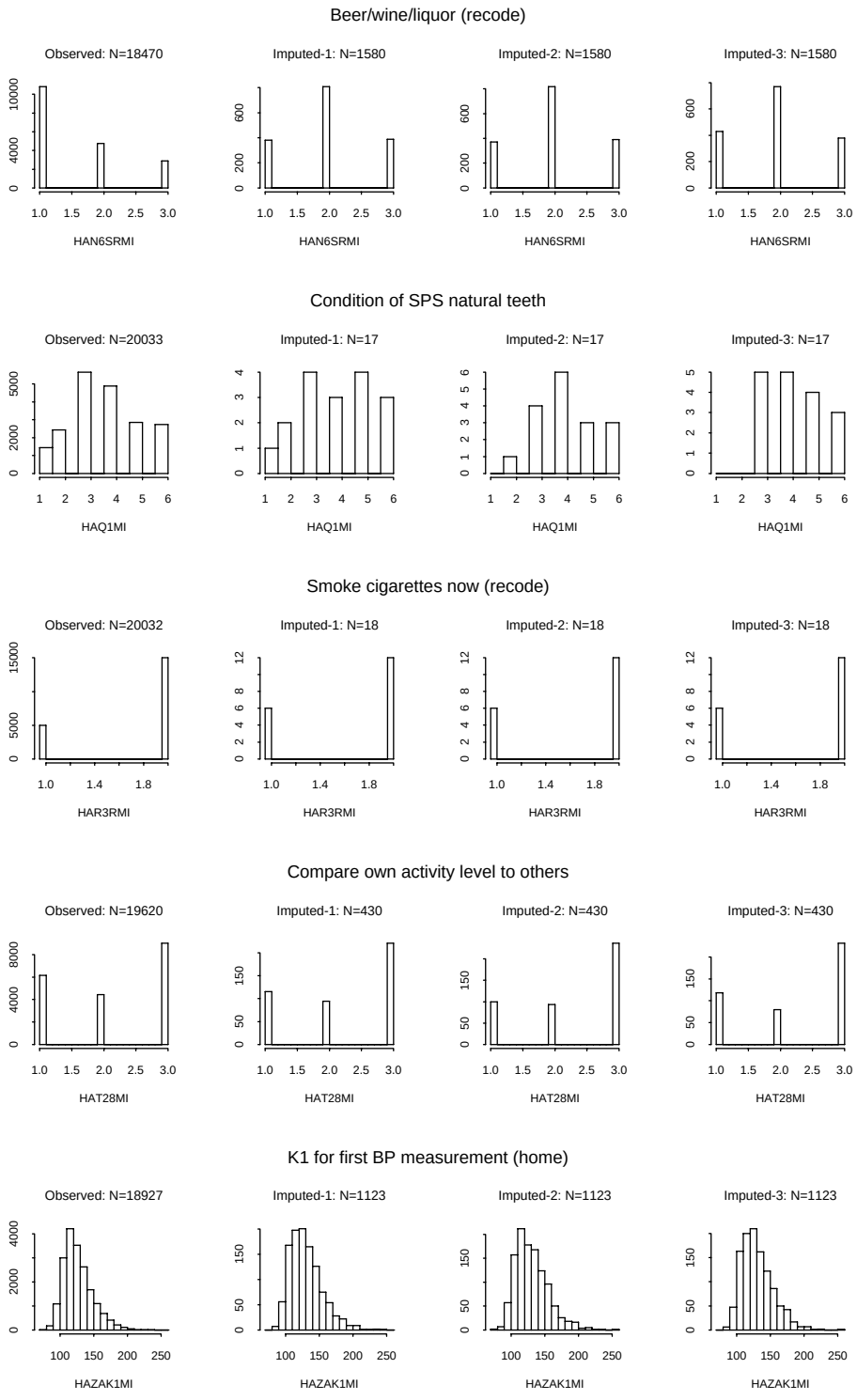

BMD of trochanter region - gm/cm sq
Observed: N=14646
Imputed-1: N=4179
Imputed-2: N=4179
Imputed-3: N=4179
BDPTRDMI
BMD of Ward's triangle region-gm/cm sq
BDPWTDMI
Buttocks circumference (cm)
Observed: N=26290
Imputed-1: N=4258
Imputed-2: N=4258
Imputed-3: N=4258
BMPBUTMI
Head circumference (cm)
Observed: N=8609
Imputed-1: N=799
Imputed-2: N=799
Imputed-3: N=799
BMPHEAMI
Standing height (cm)
Observed: N=27935
Imputed-1: N=2613
Imputed-2: N=2613
Imputed-3: N=2613
BMPHTMI

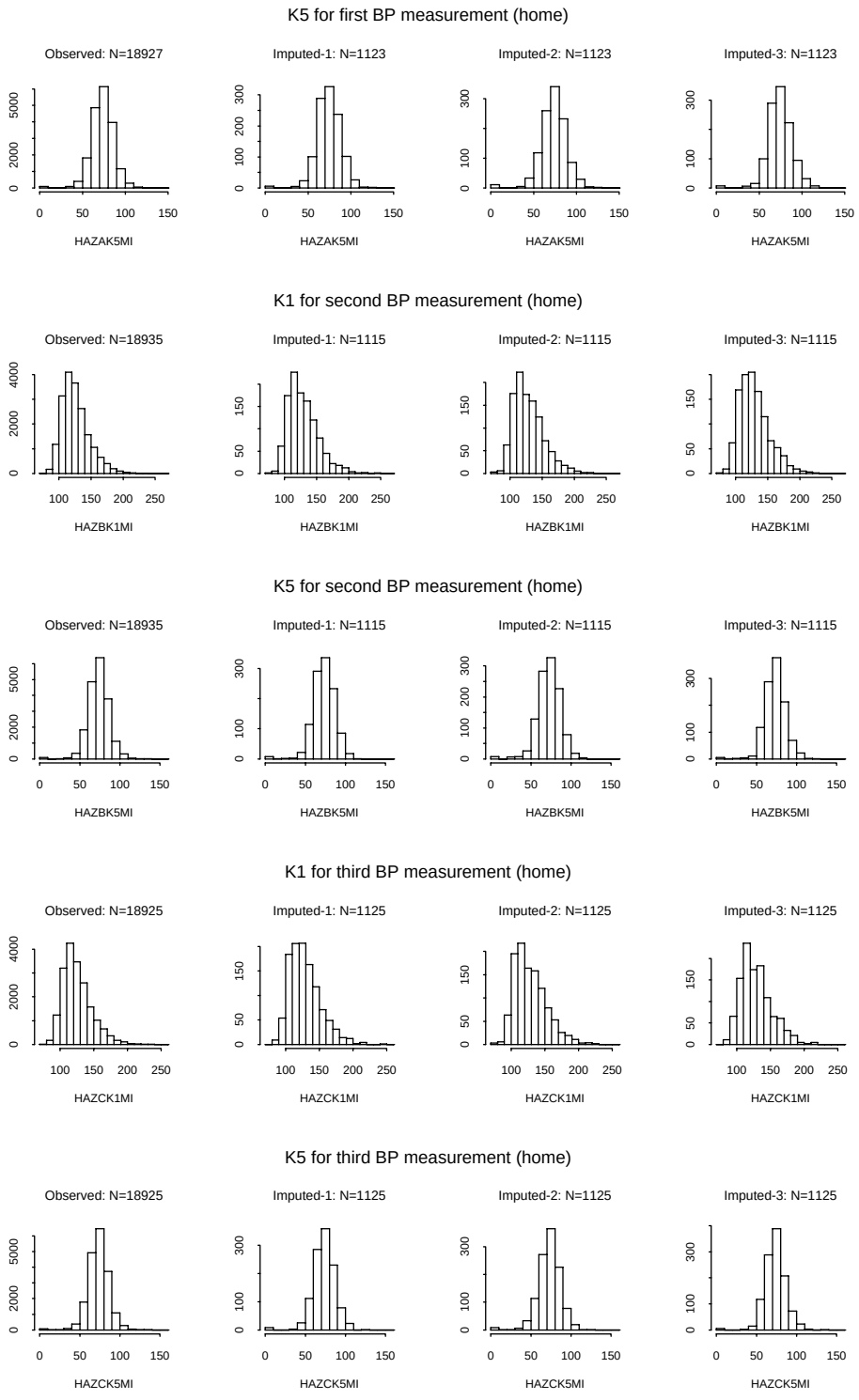

Knee height (cm)
Observed: N=4938
Imputed-1: N=1658
Imputed-2: N=1658
Imputed-3: N=1658
BMPKNEMI
Recumbent length (cm)
Observed: N=5525
Imputed-1: N=457
Imputed-2: N=457
Imputed-3: N=457
BMPRECMI
Sitting height (cm)
Observed: N=27009
Imputed-1: N=3539
Imputed-2: N=3539
Imputed-3: N=3539
BMPSTHMI
First subscapular skinfold (mm)
Observed: N=28250
Imputed-1: N=5744
Imputed-2: N=5744
Imputed-3: N=5744
BMPSB1MI
Second subscapular skinfold (mm)
Observed: N=28185
Imputed-1: N=5809
Imputed-2: N=5809
Imputed-3: N=5809
BMPSB2MI

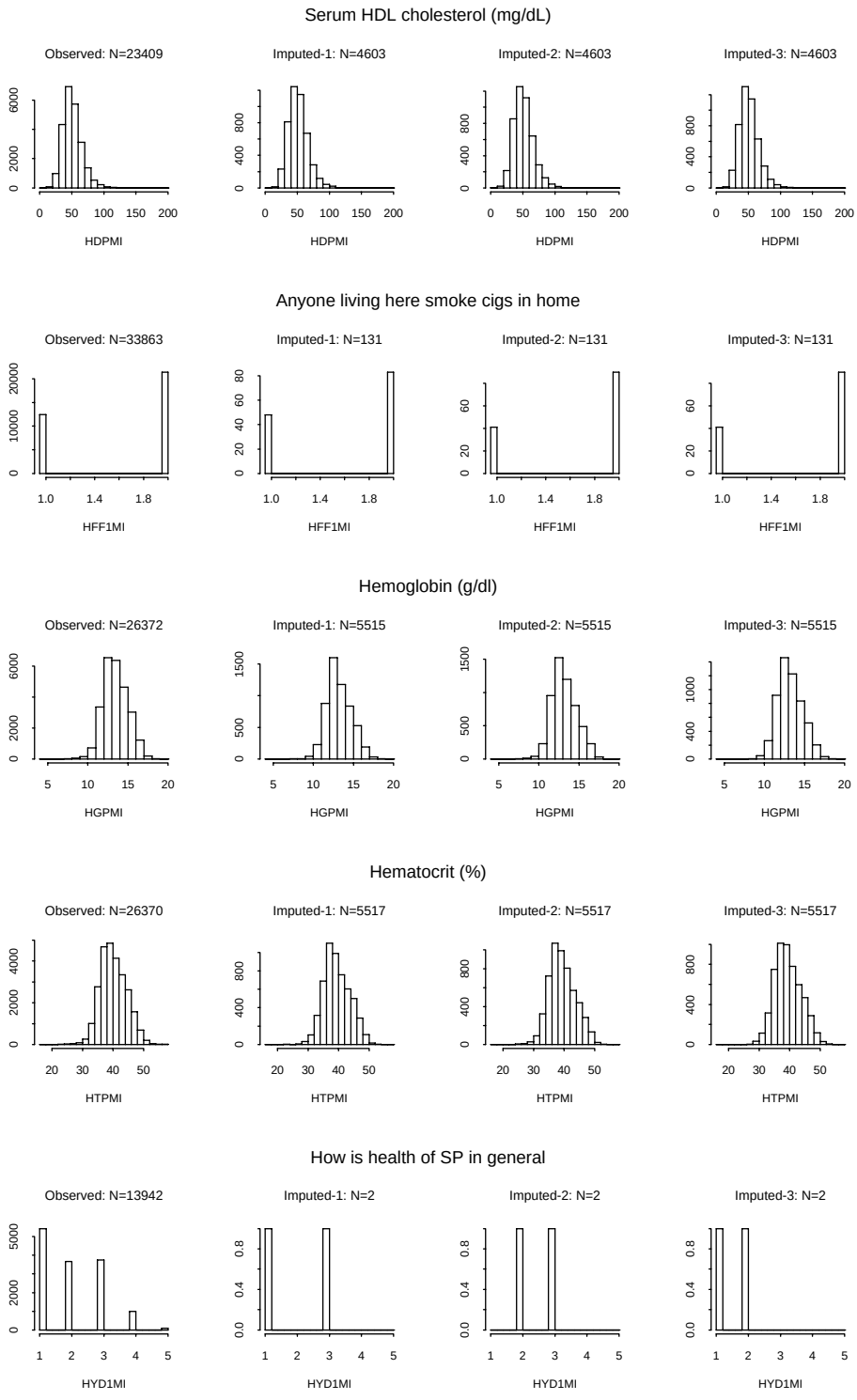

First suprailiac skinfold (mm)
Observed: N=25305
Imputed-1: N=5243
Imputed-2: N=5243
Imputed-3: N=5243
BMPSP1MI
Second suprailiac skinfold (mm)
Observed: N=25208
Imputed-1: N=5340
Imputed-2: N=5340
Imputed-3: N=5340
BMPSP2MI
First triceps skinfold (mm)
Observed: N=29394
Imputed-1: N=4600
Imputed-2: N=4600
Imputed-3: N=4600
BMPTR1MI
Second triceps skinfold (mm)
Observed: N=28886
Imputed-1: N=5108
Imputed-2: N=5108
Imputed-3: N=5108
BMPTR2MI
Waist circumference (cm)
Observed: N=26288
Imputed-1: N=4260
Imputed-2: N=4260
Imputed-3: N=4260
BMPWSTMI

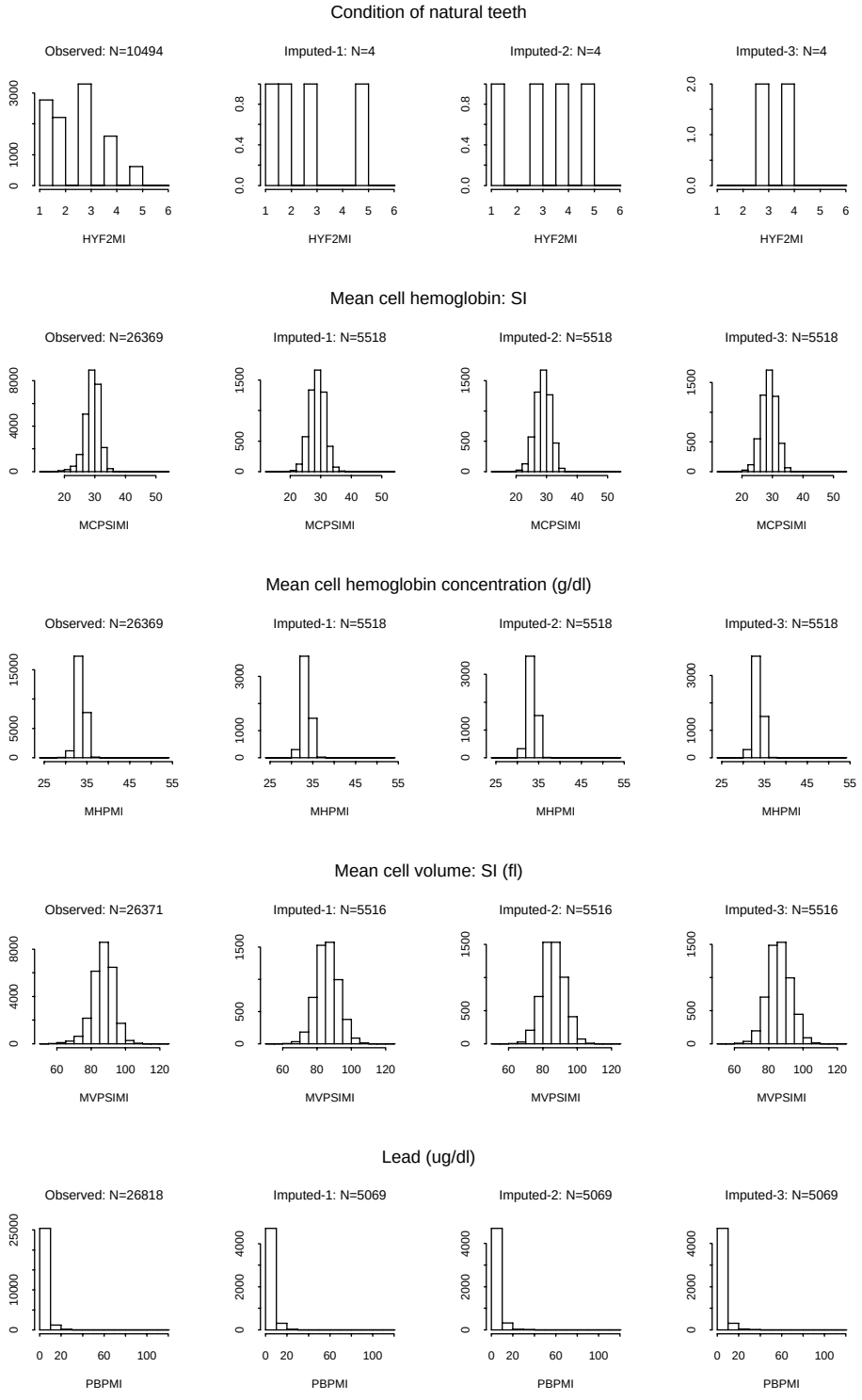

Weight (kg)
Observed: N=31132
Imputed-1: N=2862
Imputed-2: N=2862
Imputed-3: N=2862
BMPWTMI
Poverty income ratio
Observed: N=30584
Imputed-1: N=3410
Imputed-2: N=3410
Imputed-3: N=3410
DMPPIRMI
Serum iron (ug/dl)
Observed: N=26479
Imputed-1: N=5408
Imputed-2: N=5408
Imputed-3: N=5408
FEPMI
Summary drusen score
Observed: N=8486
Imputed-1: N=2962
Imputed-2: N=2962
Imputed-3: N=2962
FPPSUDMI
Summary age-related maculopathy score
Observed: N=8208
Imputed-1: N=3240
Imputed-2: N=3240
Imputed-3: N=3240
FPPSUMMI


Summary diabetic retinopathy score
Observed: N=8211
Imputed-1: N=3237
Imputed-2: N=3237
Imputed-3: N=3237
FPPSURMI
Ferritin (ng/ml)
Observed: N=26393
Imputed-1: N=5494
Imputed-2: N=5494
Imputed-3: N=5494
FRPMI
Self-rating of health status
Observed: N=20037
Imputed-1: N=13
Imputed-2: N=13
Imputed-3: N=13
HAB1MI
How tall are you without shoes-inchs
Observed: N=19129
Imputed-1: N=921
Imputed-2: N=921
Imputed-3: N=921
HAM5MI
How much do you weigh in pounds
Observed: N=19369
Imputed-1: N=681
Imputed-2: N=681
Imputed-3: N=681
HAM6MI

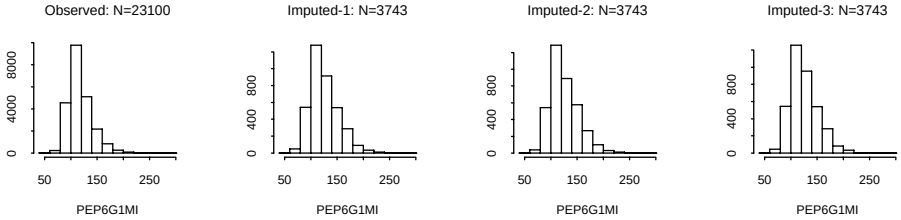

Beer/wine/liquor (recode)
Observed: N=18470
Imputed-1: N=1580
Imputed-2: N=1580
Imputed-3: N=1580
HAN6SRMI
Condition of SPS natural teeth
Observed: N=20033
Imputed-1: N=17
Imputed-2: N=17
Imputed-3: N=17
HAQ1MI
Smoke cigarettes now (recode)
Observed: N=20032
Imputed-1: N=18
Imputed-2: N=18
Imputed-3: N=18
HAR3RMI
Compare own activity level to others
Observed: N=19620
Imputed-1: N=430
Imputed-2: N=430
Imputed-3: N=430
HAT28MI
K1 for first BP measurement (home)
Observed: N=18927
Imputed-1: N=1123
Imputed-2: N=1123
Imputed-3: N=1123
HAZAK1MI


K5 for first BP measurement (home)
Observed: N=18927
Imputed-1: N=1123
Imputed-2: N=1123
Imputed-3: N=1123
HAZAK5MI
K1 for second BP measurement (home)
Observed: N=18935
Imputed-1: N=1115
Imputed-2: N=1115
Imputed-3: N=1115
HAZBK1MI
K5 for second BP measurement (home)
Observed: N=18935
Imputed-1: N=1115
Imputed-2: N=1115
Imputed-3: N=1115
HAZBK5MI
K1 for third BP measurement (home)
Observed: N=18925
Imputed-1: N=1125
Imputed-2: N=1125
Imputed-3: N=1125
HAZCK1MI
K5 for third BP measurement (home)
Observed: N=18925
Imputed-1: N=1125
Imputed-2: N=1125
Imputed-3: N=1125
HAZCK5MI

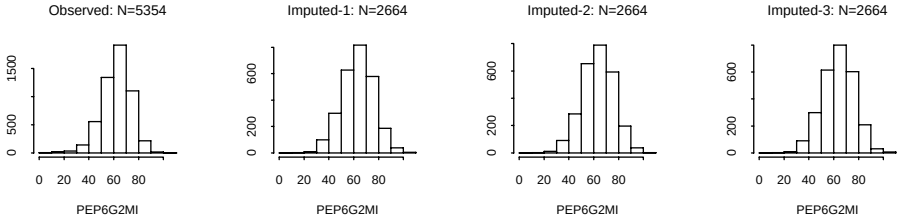

Serum HDL cholesterol (mg/dL)
Observed: N=23409
Imputed-1: N=4603
Imputed-2: N=4603
Imputed-3: N=4603
HDPMI
Anyone living here smoke cigs in home
Observed: N=33863
Imputed-1: N=131
Imputed-2: N=131
Imputed-3: N=131
HFF1MI
Hemoglobin (g/dl)
Observed: N=26372
Imputed-1: N=5515
Imputed-2: N=5515
Imputed-3: N=5515
HGPMI
Hematocrit (%)
Observed: N=26370
Imputed-1: N=5517
Imputed-2: N=5517
Imputed-3: N=5517
HTPMI
How is health of SP in general
Observed: N=13942
Imputed-1: N=2
Imputed-2: N=2
Imputed-3: N=2
HYD1MI


Condition of natural teeth
Observed: N=10494
Imputed-1: N=4
Imputed-2: N=4
Imputed-3: N=4
HYF2MI
Mean cell hemoglobin: SI
Observed: N=26369
Imputed-1: N=5518
Imputed-2: N=5518
Imputed-3: N=5518
MCPSIMI
Mean cell hemoglobin concentration (g/dl)
Observed: N=26369
Imputed-1: N=5518
Imputed-2: N=5518
Imputed-3: N=5518
MHPMI
Mean cell volume: SI (fl)
Observed: N=26371
Imputed-1: N=5516
Imputed-2: N=5516
Imputed-3: N=5516
MVPSIMI
Lead (ug/dl)
Observed: N=26818
Imputed-1: N=5069
Imputed-2: N=5069
Imputed-3: N=5069
PBPMI

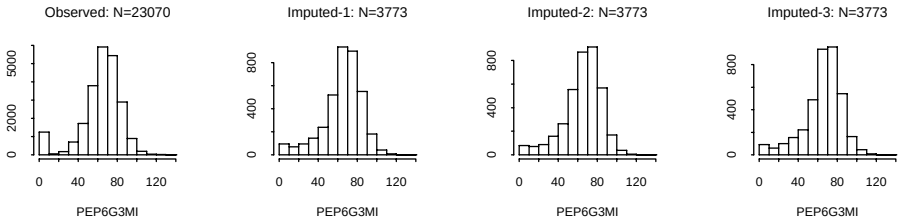
K1, systolic, for 1st BP (mmHg)



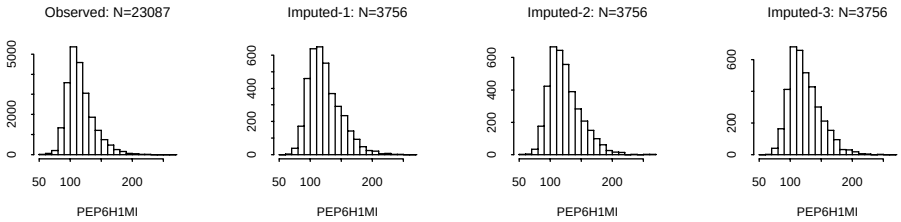
K4, diastolic, for 1st BP(mmHg)



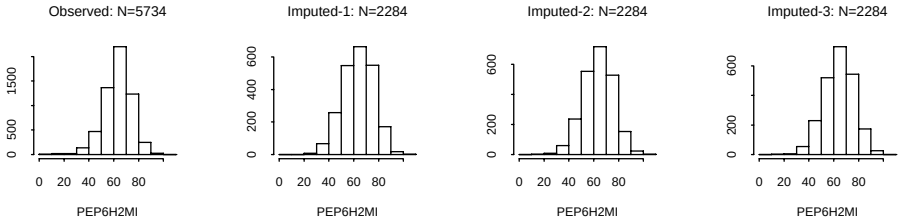
K5, diastolic, for 1st BP (mmHg)

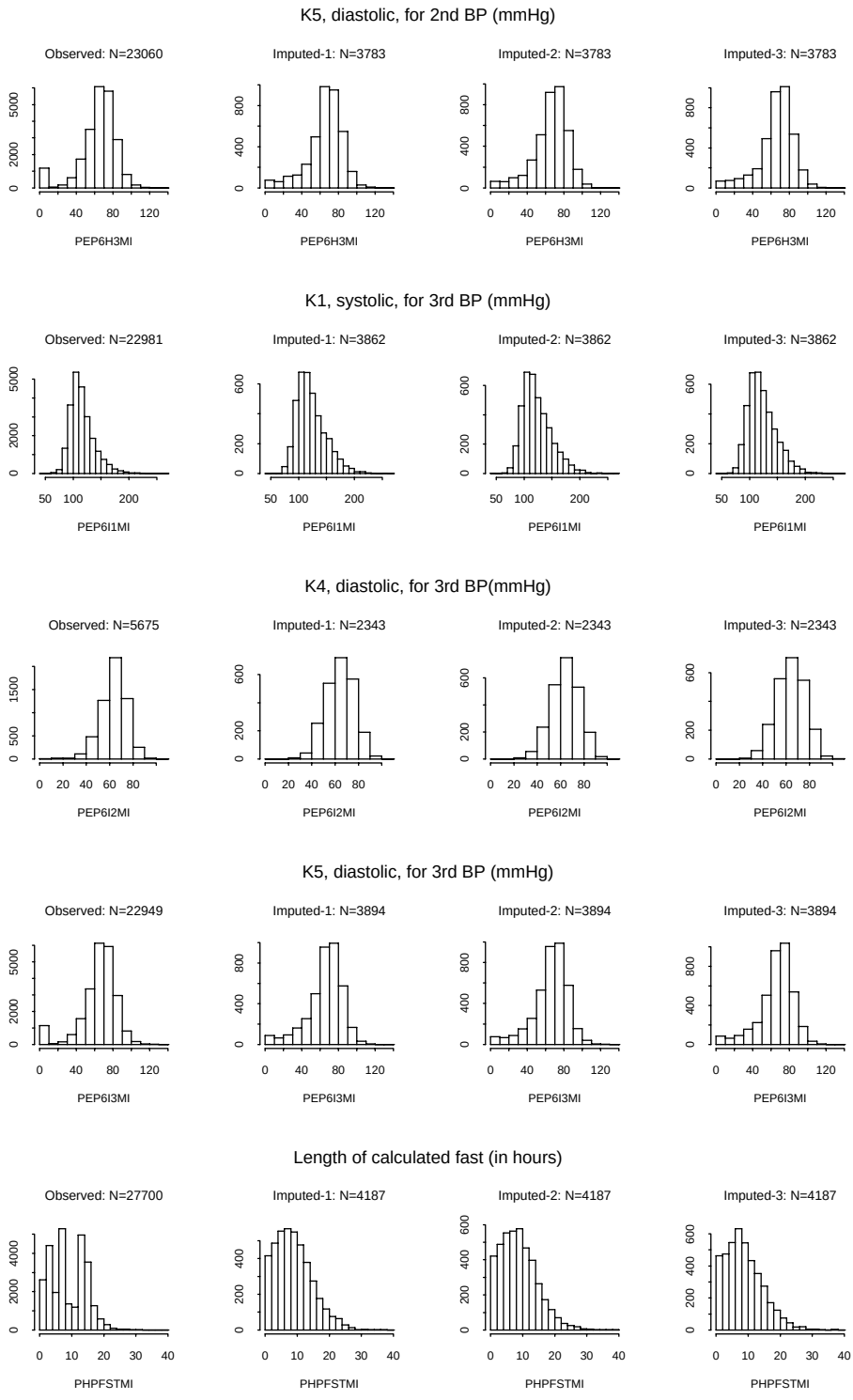


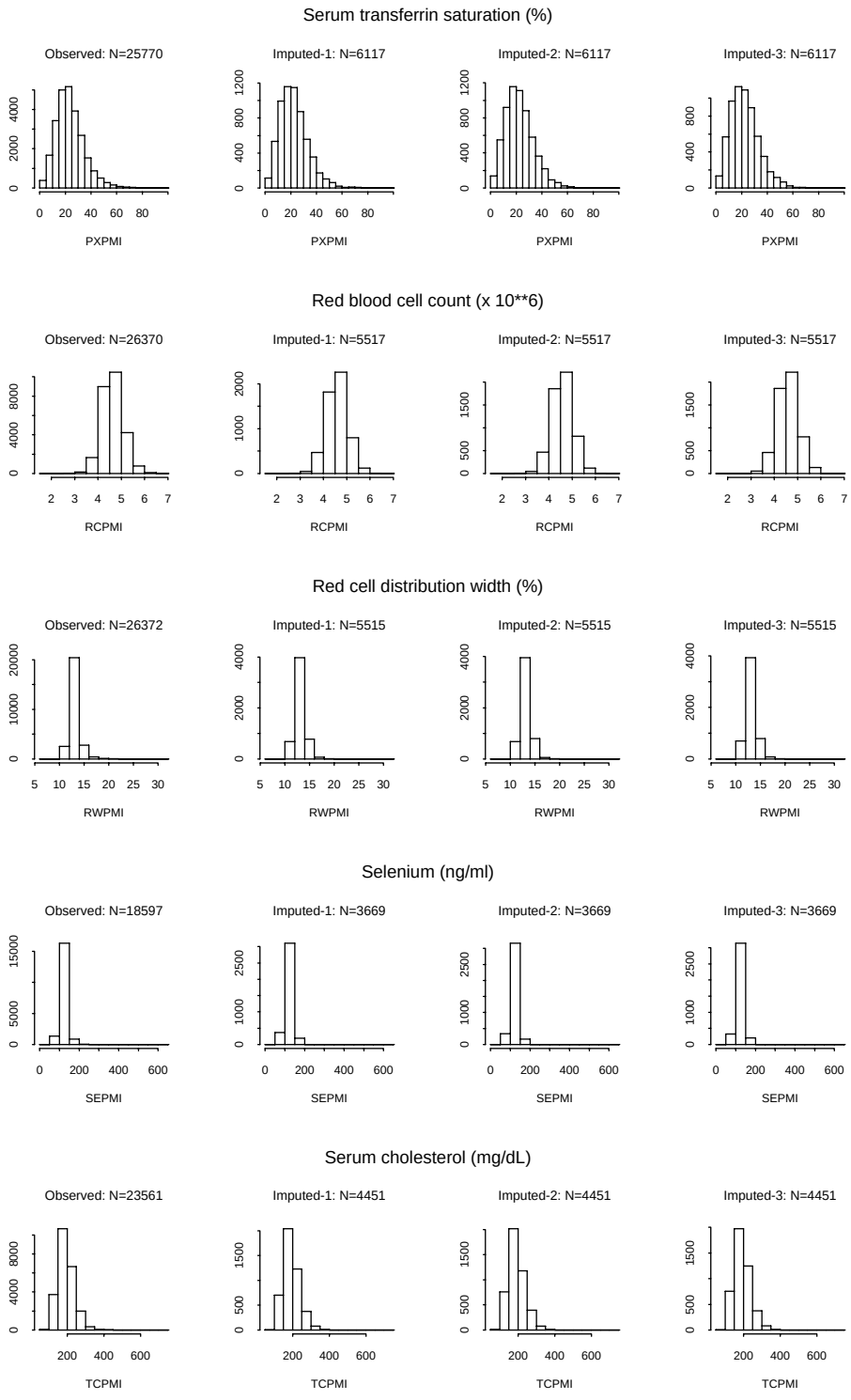
K1, systolic, for 2nd BP (mmHg)

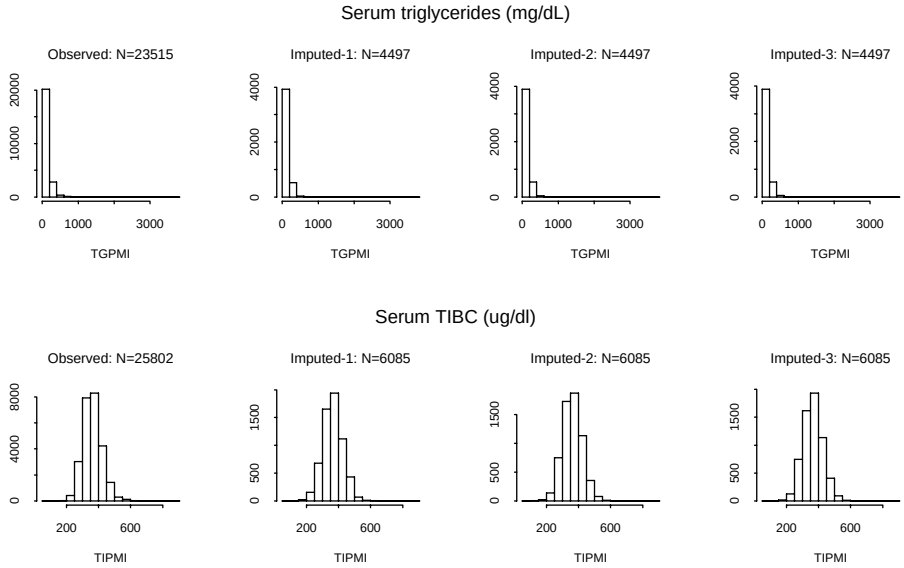


K4, diastolic, for 2nd BP(mmHg)



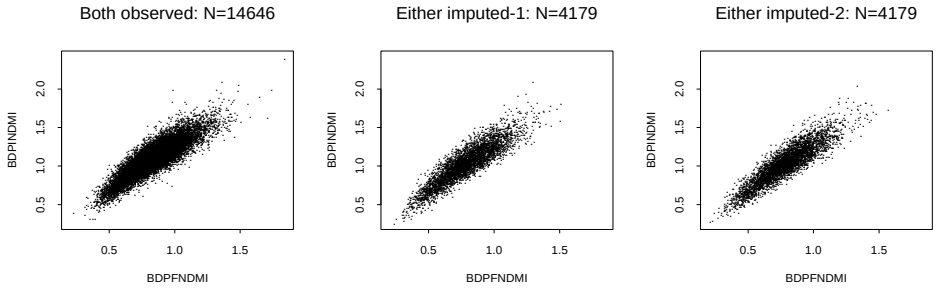

K5, diastolic, for 2nd BP (mmHg)
Observed: N=23060
Imputed-1: N=3783
Imputed-2: N=3783
Imputed-3: N=3783
PEP6H3MI
K1, systolic, for 3rd BP (mmHg)
Observed: N=22981
Imputed-1: N=3862
Imputed-2: N=3862
Imputed-3: N=3862
PEP6I1MI
K4, diastolic, for 3rd BP(mmHg)
Observed: N=5675
Imputed-1: N=2343
Imputed-2: N=2343
Imputed-3: N=2343
PEP6I2MI
K5, diastolic, for 3rd BP (mmHg)
Observed: N=22949
Imputed-1: N=3894
Imputed-2: N=3894
Imputed-3: N=3894
PEP6I3MI
Length of calculated fast (in hours)
Observed: N=27700
Imputed-1: N=4187
Imputed-2: N=4187
Imputed-3: N=4187
PHPFSTMI


Serum transferrin saturation (%)
Observed: N=25770
Imputed-1: N=6117
Imputed-2: N=6117
Imputed-3: N=6117
PXPMI
Red blood cell count (x 10**6)
Observed: N=26370
Imputed-1: N=5517
Imputed-2: N=5517
Imputed-3: N=5517
RCPMI
Red cell distribution width (%)
Observed: N=26372
Imputed-1: N=5515
Imputed-2: N=5515
Imputed-3: N=5515
RWPMI
Selenium (ng/ml)
Observed: N=18597
Imputed-1: N=3669
Imputed-2: N=3669
Imputed-3: N=3669
SEPMI
Serum cholesterol (mg/dL)
Observed: N=23561
Imputed-1: N=4451
Imputed-2: N=4451
Imputed-3: N=4451
TCPMI

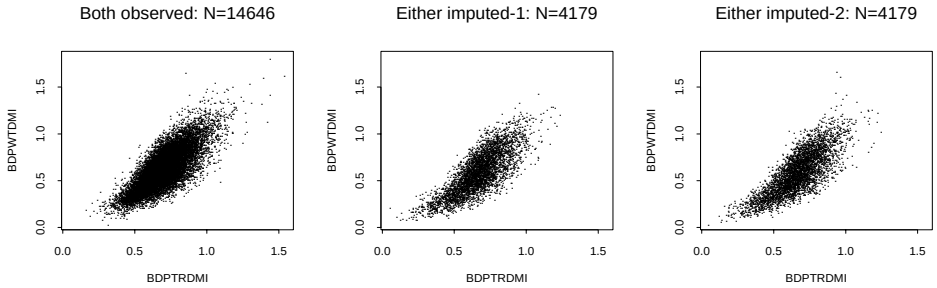

Serum triglycerides (mg/dL)
Observed: N=23515
Imputed-1: N=4497
Imputed-2: N=4497
Imputed-3: N=4497
0
10000
20000
2000
4000
0
1000
3000
TGPMI
Serum TIBC (ug/dl)
Observed: N=25802
Imputed-1: N=6085
Imputed-2: N=6085
Imputed-3: N=6085
0
4000
8000
500
1500
200
600
TIPMI

Appendix C: Bivariate comparisons

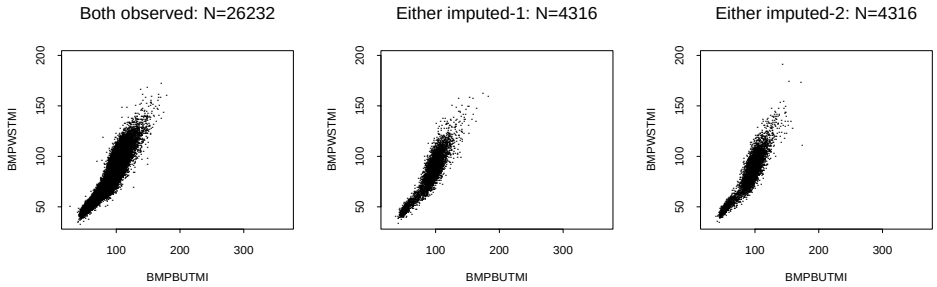
BMD of femur neck versus intertrochanter region



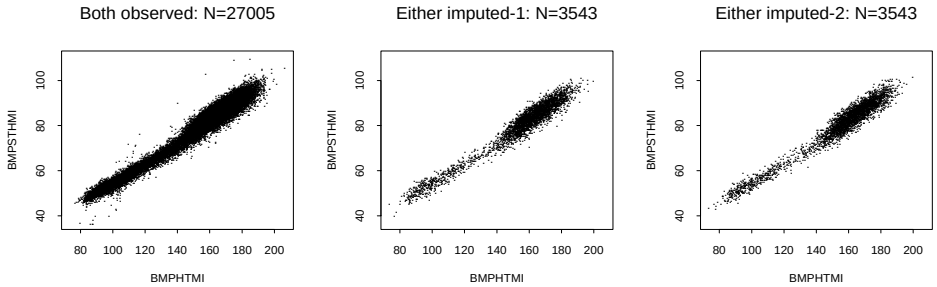
BMD of trochanter versus Ward's triangle region

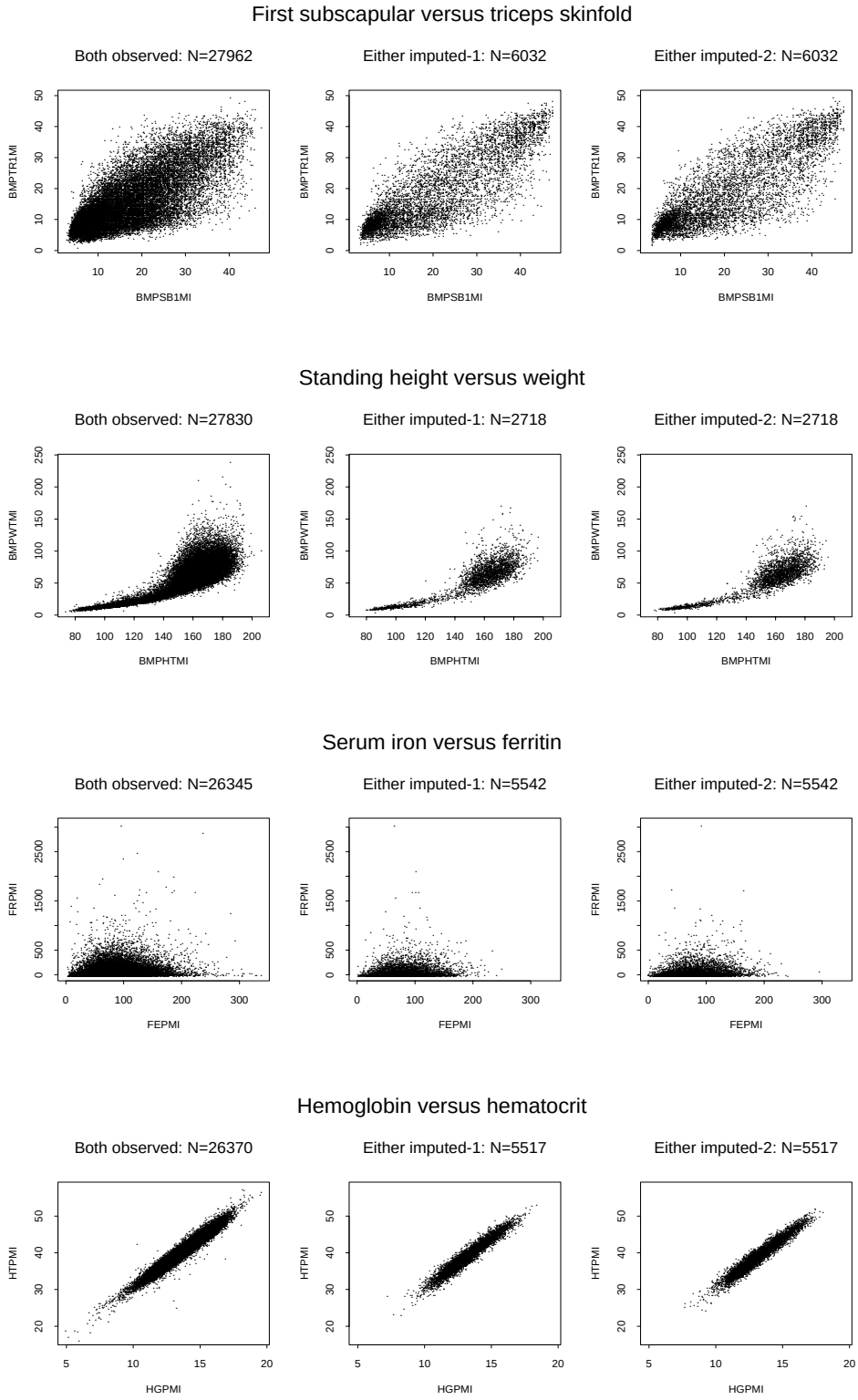


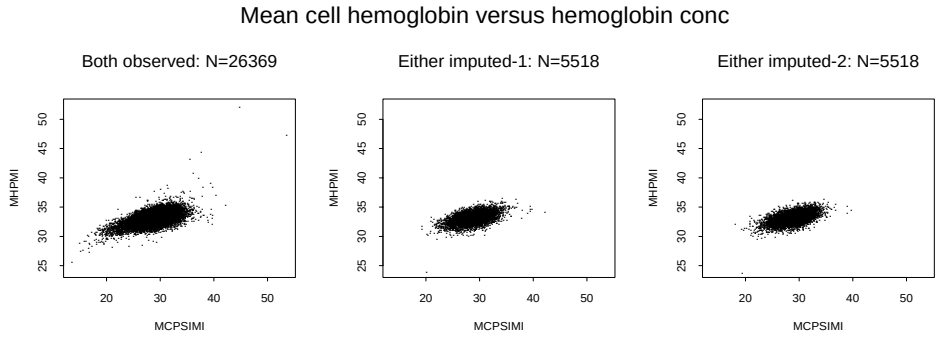
Buttocks versus waist circumference

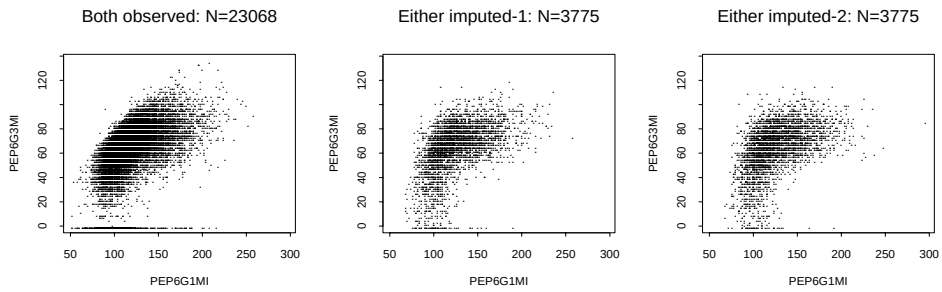


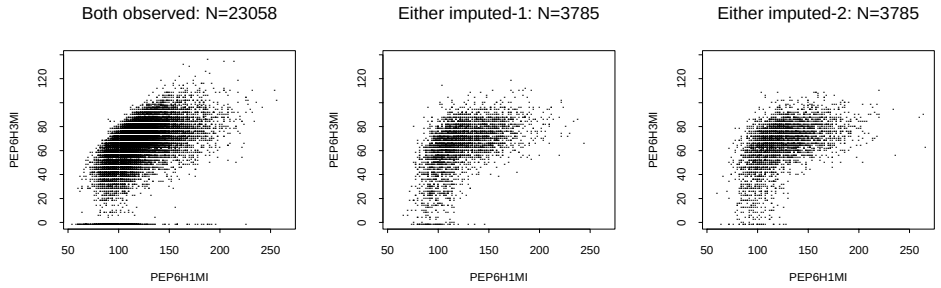
Standing height versus sitting height

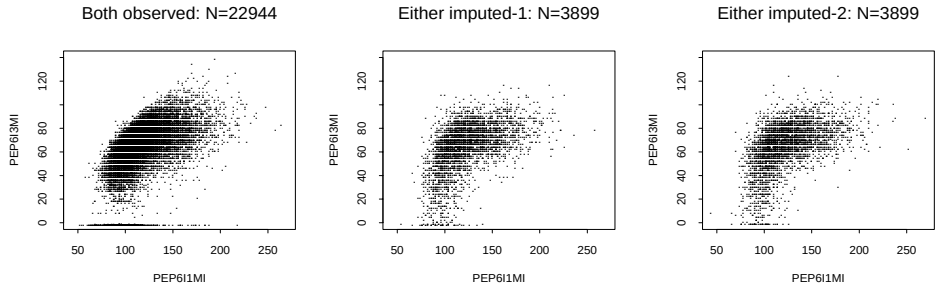


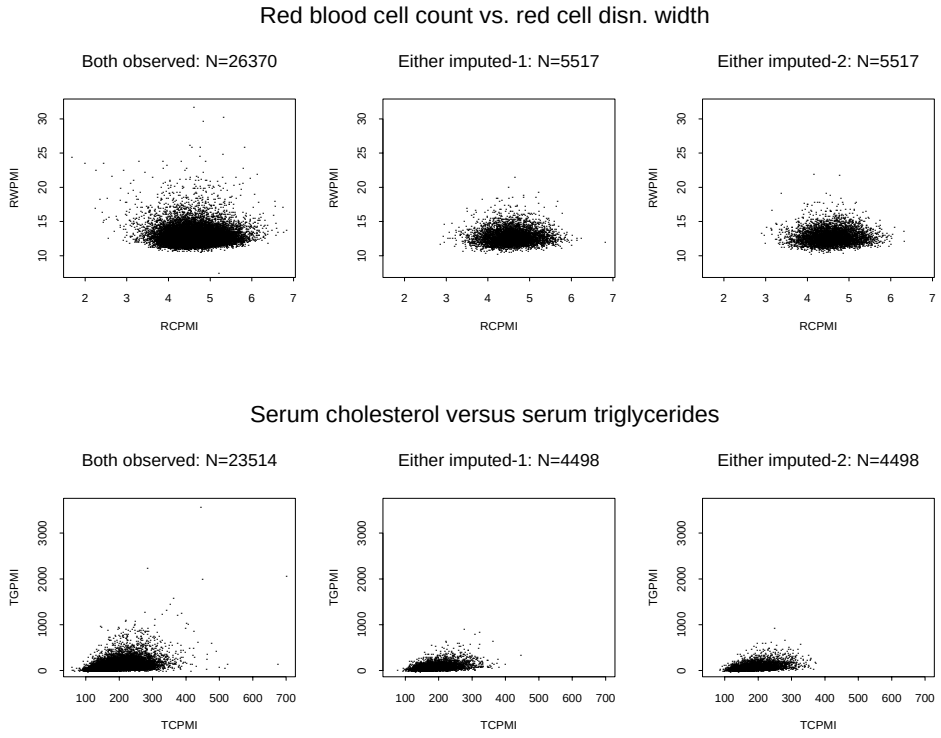

First subscapular versus triceps skinfold
Both observed: N=27962
Either imputed-1: N=6032
Either imputed-2: N=6032
BMPTR1MI
BMPSB1MI
Standing height versus weight
Both observed: N=27830
Either imputed-1: N=2718
Either imputed-2: N=2718
BMPWTMI
BMPHTMI
Serum iron versus ferritin
Both observed: N=26345
Either imputed-1: N=5542
Either imputed-2: N=5542
FRPMI
FEPMI
Hemoglobin versus hematocrit
Both observed: N=26370
Either imputed-1: N=5517
Either imputed-2: N=5517
HTPMI
HGPMI


Mean cell hemoglobin versus hemoglobin conc
Both observed: N=26369
Either imputed-1: N=5518
Either imputed-2: N=5518
MHPMI
MCPSIMI


K1 systolic (1) versus K5 diatolic (1)
Both observed: N=23068
Either imputed-1: N=3775
Either imputed-2: N=3775
PEP6G3MI
PEP6G1MI


K1 systolic (2) versus K5 diatolic (2)
Both observed: N=23058
Either imputed-1: N=3785
Either imputed-2: N=3785
PEP6H3MI
PEP6H1MI


K1 systolic (3) versus K5 diatolic (3)
Both observed: N=22944
Either imputed-1: N=3899
Either imputed-2: N=3899
PEP6I3MI
PEP6I1MI


Red blood cell count vs. red cell disn. width
Both observed: N=26370
Either imputed-1: N=5517
Either imputed-2: N=5517
RWPMI
RCPMI
Serum cholesterol versus serum triglycerides
Both observed: N=23514
Either imputed-1: N=4498
Either imputed-2: N=4498
TGPMI
TCPMI