STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 38
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27-29 May 2002, Helsinki, Finland)

Topic (iv): Impact of new technologies on statistical data editing

USE OF DATA EDITING AND MULTIPLE IMPUTATION IN HEALTH SURVEYS

Contributed paper

Submitted by National Center for Health Statistics, CDC/DHHS, United States[1]

Abstract: The National Center for Health Statistics (NCHS) is the Federal agency responsible for the collection and dissemination of the nation's vital and health statistics. To carry out its mission, NCHS conducts a wide range of annual, periodic and longitudinal sample surveys and administers the national vital statistics registration systems. Over the years, NCHS has developed a number of mechanisms to assure that the products of its sample surveys and registration systems meet standards of completeness, reliability and validity. Two such mechanisms are data editing and imputation. This paper will describe some of the most frequently used data editing procedures in NCHS health surveys and will illustrate the use of multiple imputation in the National Health and Nutrition Examination Survey (NHANES), which is NCHS' most complex survey.

I. INTRODUCTION

1. The National Center for Health Statistics (NCHS) is the Federal agency responsible for the collection and dissemination of the nation's vital and health statistics. To carry out its mission, NCHS conducts a wide range of annual, periodic, and longitudinal sample surveys and administers the national vital statistics registration systems. These sample surveys and registration systems form four families of data systems: vital event registration systems, population based surveys, provider based surveys, and followup/followback surveys.

2. Much of what happens to the data covered by these data systems, from collection through publication, depends on the family to which they belong. At most steps along the way, various activities and operations are implemented with the goal of making the data as accurate as possible. These activities and operations are generally categorized under the rubric, "data editing." In the 1990 Statistical Policy Working Paper 18, "Data Editing in Federal Statistical Agencies," [1] data editing is defined as:

---

[1] Prepared by Kenneth W. Harris (KWH1@CDC.GOV)

3. Procedure (s) designed and used for detecting erroneous and/or questionable survey data (survey response data or identification type data) with the goal of correcting (manually and/or via electronic means) as much of the erroneous data (not necessarily all of the questioned data) as possible, usually prior to data imputation and summary procedures.

4. Detailed descriptions of NCHS data editing practices and procedures covering more than 20 data systems were provided in an earlier report [2]. This paper provides summary data editing procedures in the following areas:

- Environment in Which the Survey Takes Place
- Data Processing Environment and Dispersion of the Work
- Micro-, Macro-, and Statistical Editing
- Prioritizing of Edits
- Imputation Procedures
- Editing and Imputation Standards
- Role of Subject Matter Specialists
- Measures of Variation

5. Within each of these areas, data editing practices are grouped according to the type of data system, i.e., vital event registration systems, population based surveys, provider based surveys, and followup/followback surveys.

6. In addition, the paper will discuss multiple imputation methodology and how it is used in the National Health and Nutrition Examination Survey (NHANES), which is NCHS' most complex survey.

II. ENVIRONMENT IN WHICH THE SURVEY TAKES PLACE

Registration Systems

7. The vital event registration systems cover five vital events: mortality, fetal mortality, natality, marriage and divorce. For each of these systems, data are obtained in electronic form from state registration offices and registration offices of selected cities and other areas. Coverage for each registration system is limited to its prescribed registration area (RA). The oldest registration areas, mortality, fetal mortality, and natality, were complete by 1933. These are national data systems; i.e., they cover the entire United States. The marriage RA started in 1957 with 30 states and reached its current coverage of 42 states plus selected areas in 1986. The Divorce RA started in 1958 with 14 states and by 1986 had expanded to 31 states plus selected areas [3,4].

8. Mortality (approximately 2,000,000 annual events), fetal mortality (60,000) and natality (4,000,000) registration are required by all states; registration completeness for the mortality and natality systems exceeds 99 percent. The Marriage RA, excluding Puerto Rico and the Virgin

Islands, covers approximately 81 percent (785,000) of U.S. marriages. The Divorce RA, excluding the Virgin Islands, accounts for 49 percent (280,000) of the annual U.S. divorce count.

9. In addition to these five registration systems, one other data system, the Linked Birth and Infant Death Data Set, is based on data obtained from the Mortality and Natality Registration Systems. The Linked Birth and Infant Death Data Set, which also covers the 50 states, the District of Columbia and New York City, links the more detailed information from the birth certificate with the information from the death certificate for each of the approximately 40,000 infants who dies before his/her first birthday.

Population Based Surveys

10. Five of the Center's data systems are classified as population based surveys. They are the National Health Interview Survey (NHIS), National Health and Nutrition Examination Survey (NHANES), the National Survey of Family Growth (NSFG), the National Immunization Survey (NIS), and the State and Local Area Integrated Telephone Survey (SLAITS). The designs of these surveys are based on stratified multistage samples of households, where the household is defined as the basic sample unit. Based on established criteria, a person (one or more) in the sample household is selected as the ultimate sample unit, i.e., the unit of analysis.

National Health Interview Survey (NHIS)

11. The **NHIS** is a continuing nationwide sample survey [5] in which data are collected on the incidence of acute illness and injuries, the prevalence of chronic conditions and impairments, the extent of disability, the utilization of health care services, and other health related topics. Generally, personal interviews are completed in 40,000 households for about 100,000 sample persons.

National Health and Nutrition Examination Survey (NHANES)

12. The **NHANES** obtains nationally representative information on the health and nutritional status of the American population through a combination of personal interviews (mostly in the respondent's home) and detailed physical examinations. These examinations are conducted in specially equipped mobile examination centers (MECs) that travel around the country. The latest survey, NHANES III, the sixth in the cycle of health examination surveys conducted since 1960 [6], collected data on topics such as high blood pressure, blood cholesterol, infectious diseases, diabetes, HIV infection, blood lead levels, allergies, osteoporosis, and other nutritional status measures.

13. The **NHANES** III [7], conducted over two 3 year phases, 1988-91 and 1991-94, covered the U.S. civilian, noninstitutional population aged 2 months and older. Each phase constituted a national sample of about 20,000 persons, with an expected interview completion rate of 85-90 percent and a response rate of about 75-80 percent for the medical examination. More than 78

percent of the persons selected for the 1988-91 phase participated in the medical examination. Selected subpopulations, children (< 5 years), older persons (60+), Black Americans and Mexican Americans, were oversampled.

National Survey of Family Growth (NSFG)

14. The **NSFG** is a periodic nationally representative household survey of women of reproductive age (15-44 years). The survey, first conducted in 1973 [8], collects data on fertility and infertility, family planning, and related aspects of maternal and infant health. (Sample sizes ranged from 7,969 to 10,847.) Five cycles of the survey were conducted between 1973 and 1995. The sixth survey, scheduled for 2002, will also include men in its coverage.

National Immunization Survey (NIS)

15. The NIS, whose target population is children 19 to 35 months of age living in the United States (at the time of interview) collects specific vaccination information that allows coverage rates to be monitored at national, state, and local area levels [9]. Official coverage estimates reported from the NIS give rates of being up-to-date (UTD) with respect to the recommended numbers of doses of six vaccines. These vaccines and their recommended number of doses are: diphtheria and tetanus toxoids and pertussis vaccine (DTP) - 4 doses; poliovirus vaccine (polio) - 3 doses; measles-containing vaccine (MCV) - 1 dose; *Haemophilus influenzae* type b vaccine (Hib) - 3 doses; hepatitis B vaccine (Hep B) - 3 doses; and varicella zoster vaccine (varicella) - 1 dose. In addition to these vaccines, interest focuses on coverage rates for series of vaccines, including the 4:3:1:3 series (4 DTP, 3 polio, 1 MCV, and 3 Hib).

State and Local Area Integrated Telephone Survey (SLAITS)

16. The SLAITS was designed as a household telephone interview survey on a varying set of health and welfare related topics [10]. The content and the sample design for each SLAITS module are customized to meet the needs of survey sponsors. Survey sponsors have included health agencies and health care provider associations, such as the Maternal and Child Health Bureau and the American Academy of Pediatrics. The modules requested by these sponsors address current health issues and interests, and generally are designed to describe the health, health care, or welfare of a particular population at one point in time.

Provider Based Surveys

17. Five NCHS data systems form the family of provider based surveys, collectively called the National Health Care Survey (NHCS). Included here are the National Hospital Discharge Survey (NHDS), National Hospital Ambulatory Medical Care Survey (NHAMCS), National Ambulatory Medical Care Survey (NAMCS), National Nursing Home Survey (NNHS), and the National Home and Hospice Care Survey (NHHCS). Whereas population based surveys use the household as the basic sample unit, provider based surveys use the medical provider (physician,

hospital, nursing home, etc.) as the basic sample unit. The provider furnishes information on samples of provider /patient contacts, e.g., office visits, hospital stays, nursing home stays, etc.

18. Samples for these surveys range in size from the approximately 475 emergency rooms in the NHAMCS to the 2,500 physicians covered by the NAMCS.

Followup/Followback Surveys

19. The Second Longitudinal Study of Aging, LSOA II, is the only currently "active" followup/followback survey. Three others, the National Health and Nutrition Examination Survey Epidemiologic Followup Study (NHEFS), Longitudinal Study of Aging I (LSOA), and the National Nursing Home Survey Followup (NNHSF), are included because analyses are still being conducted and/or additional waves may be conducted in the future.

III. DATA PROCESSING ENVIRONMENT AND DISPERSION OF THE WORK

20. There are many similarities in the data processing activities employed by NCHS offices for their respective data systems. This is especially true for data systems within the same "family of surveys." For example, registration areas provide NCHS with coded and edited electronic form (CDs, diskettes, etc) copies of vital event certificates which are converted to uniform codes and subjected to machine edits. The other surveys use CAPI (Computer Assisted Personal Interview), preliminary hand edits, machine edits, etc. There are, however, a number of procedures that cross "family survey" lines that are gaining greater usage with the rapid advances made in survey technology. Of particular interest to NCHS is "source point data editing" (SPDE). This refers to editing survey data by any means of access to either the interviewer (or other data collector), the respondent, or records within a limited time following the original interview or data collection. The time limit reflects the period within which the persons involved can reasonably be expected to remember details of the specific interview or, in the case of data collected from records, a time within which there is reasonable expectation that there has been no change to the records which would affect the data collected. Thus, data completion and accuracy are much more likely to result when source point data editing is used.

Audit Trail

21. This term refers to a process of maintaining, either by paper or electronically, an accounting of all changes of sample or survey data item values and the reasons for those changes. The level of effort varies by data systems; some are manual, while others are automated.

IV. MICRO-, MACRO-, AND STATISTICAL EDITING

22. This section describes three types of editing processes. The following definitions are used in this section.

- Micro-editing: Editing done at the record or questionnaire level.

- Macro-editing: Editing to detect individual errors by checking on aggregated data or by applying checks to the complete set of records. These checks are usually based on the impact the errors (would) have on the estimates [11].

- Statistical editing: A set of checks based on statistical analysis of respondent data, e.g., the ratio of two fields lies between limits determined by a statistical analysis of that ratio for presumed valid reporters [12]. A statistical edit may incorporate cross-record checks, e.g., the comparison of the value of an item in one record against a frequency distribution for that item for all records. A statistical edit may also use historical data on a firm by firm basis in a time series modeling procedure.

23. Micro-, macro, and statistical editing for the registration systems are all very similar. Automated edits are designed to (1) assure code validity for each variable and (2) verify codes or code combinations which are considered either impossible or unlikely occurrences.

24. For each of the other three types of data systems, most or all of the following procedures are used:

- Extensive machine micro-editing.
- Where appropriate, comparison of current estimates with previous years.
- Assuring reasonableness of record counts, sampling rates, etc.
- Checking ranges, skip patterns, consistency of data from different sources.
- Checking medical data for compatibility with age and/or sex.

V. PRIORITY OF EDITS

25. None of the registration systems gives special priority to any item in the editing procedures. The other data systems prioritize their edits based on:

- Identifiers needed to link data files.
- Questionnaire items used to weight sample data to national estimates.
- Medical data incompatible with demographic data.

VI. IMPUTATION PROCEDURES

26. Imputation is defined as a process for entering a value for a specific data item where the response is missing or unusable.

Registration Systems

27. Imputation procedures among registration systems apply primarily to demographic items. In Mortality registration, imputation procedures are done by machine, which checks for invalid codes. The following variables are subject to imputation procedures: age, sex, date of death, marital status of decedent, race of decedent, and education of decedent.

28. Missing natality data that are imputed include child's race, sex, date of birth, and plurality. Data imputed for the mother include race, age, marital status and residence. Imputation is done by machine, either on the basis of a previous record with similar information for other items on the record (e.g., mother's age imputed on the basis of a previous record with the same race and total-birth order), or on the basis of other information on the certificate (e.g., marital status on the basis of mother's and father's names, or lack of name). The tape documentation includes flags to indicate when imputation was performed.

29. Finally, marriage and divorce data imputation are limited to month of marriage (or divorce) and age of bride and/or groom (marriage only). Hot deck and cold deck imputation procedures are used. In hot deck imputation, a missing data item is assigned the value from a preceding record in the same survey having similar (defined) characteristics. In cold deck imputation, a missing data item is assigned the value from a similar record in a previous similar survey.

Population Based Surveys

30. Imputation procedures for the Center's other surveys differ from those used by the Registration Systems. In the case of the NHIS, unit nonresponse (missing sample cases) is imputed by inflating the sample case weight by the reciprocal of the response rate at the final stage of sample selection, or by a poststratification adjustment based on independent estimates of the population size in 88 age-race-sex/ethnicity categories.

31. Item non-response (missing question answers) is imputed, where possible, by inferring a certain or probable value from existing information for the respondent or household. For example, in the NHIS, which imputes for the above categories only, if the respondent does not know the race/ethnicity of a household member, that person is assigned to a category according to a hierachical scale beginning with the race/ethnicity of family members, other household members, neighborhood, etc. Prior to 2000, these missing items were assigned survey modal values.

32. Non-response for the income item on the NHIS is currently around 25 percent. NHIS is currently conducting a feasibility study testing the applicability of a sequential regression multiple imputation procedure developed by the University of Michigan. The procedure is being tested on 2 variables, family income and personal earnings.

33. In the NHANES, the calculation of sample weights addresses the unit nonresponse aspects of the survey except for special cases. The institution of multiple imputation procedures in NHANES III will be discussed in a later section of this paper.

34. In the NSFG, the sample weights adjust for unit nonresponse. Imputation of missing items in the last NSFG was carried out by the contractor. For the most part, a hot-deck procedure was used to impute missing values.

Provider Based Surveys

35. The provider based surveys have established imputation procedures for three types of nonresponse: unit nonresponse, record nonresponse, and item nonresponse. Unit nonresponse is imputed by inflating the sample weight of similar responding units. Record nonresponse is imputed by inflating the sample weight of similar responding cases to account for the missing cases. Item nonresponse is imputed by inferring a certain or probable value from existing respondent information. For example, beginning with the 1996 NHDS, a hot deck for missing values of age and sex has been used that maintains the known age or sex distribution of records within the same 3-digit level of first-listed diagnostic code.

Followup/Followback Surveys

36. The followback surveys did not impute any data, although selected missing data items were filled in by using logical relationships with certain completed items. For example, in LSOA and LSOA II, if the item on “Retirement Status” was blank but the item on “Source of Retirement Income” was completed, then “Retirement Status” could logically be completed. Unknown or inconsistent data were coded as “unknown.”

VII. EDITING AND IMPUTATION STANDARDS

37. For each of its registration systems, NCHS monitors the quality of demographic and medical data received from the states (tapes, electronically) by independent verification of a sample of records of data entry errors. In addition, there is verification of coding at the state level before NCHS receives the data. All other systems employ error tolerance standards established for interviewer performance (if applicable), and enforced by editing and telephone reinterviews.. Error tolerance standards are also established for coding and keying of data, and are enforced by sample verification.

VIII. ROLE OF SUBJECT MATTER SPECIALISTS

38. For all surveys and data systems, the primary role of subject matter specialists is to write edit specifications (from which edit programs are prepared) to review results of edit runs and to adjudicate failures in collaboration with programmers. Their secondary role is to compare standard sets of estimates with historical series to identify anomalies. In addition, they also consult with survey design staff on field edits.

IX. MEASURES OF VARIATION

Registration Systems

39. For the four vital event registration systems that conduct 100 percent surveys, i.e., Mortality, Fetal Mortality, Natality, and Linked Birth and Infant Death, there is no sampling error. However, for comparisons of small numbers of vital events (e.g., less than 20 deaths), the computation of estimates of variation is outlined in two Technical Appendixes [11,12].

Population Based Surveys

40. Estimates of sampling errors are produced and published for each survey. No estimates of nonsampling error are produced. Additionally, since the NSFG sample was drawn from households with known characteristics (because they had participated in the NHIS), some estimates of nonresponse error were also made, and sampling weights were corrected to try to compensate for those errors.

41. For the NHANES, an entire series of publications, aimed solely at methodological issues, disseminates information on all aspects of survey design and sampling concerns. For example, one publication deals with regression techniques, analysis of variance and categorical data analysis [13].

42. For all surveys, estimates of sampling error are produced and published for each data year. The NHDS also uses specified guidelines for presenting estimates based on assumptions regarding the probability distribution of the sampling error. No estimates of non sampling error are produced.

Followup/Followback Surveys

43. Estimates of sampling errors are included in reports describing survey results.

X. MULTIPLE IMPUTATION

44. As described earlier, most NCHS surveys rely on long established single imputation techniques for addressing non- response at the unit, record, or item level. In the NHANES , for example, unit non-response was generally addressed through the calculation of sample weights.

45. These various single imputation techniques, though providing estimates for inclusion in NCHS surveys, are known to have limitations.

46. Over the last decade or so, many surveys have experienced increased non-response rates. This is especially so for large multi-stage surveys such as the NHANES.

47. The NHANES III, first conducted in 1960 as the National Health Examination Survey (NHES), is by far the most challenging of the Center's surveys , encompassing a complex survey design that includes both a home interview section and a subsequent detailed medical examination conducted in specially equipped Mobile Examination Centers (MECs).

48. The NHANES was designed as a six year survey to be conducted in two phases, 1988-1991 and 1991-1994. At the outset, it was expected to yield an overall response rate (i.e., both interview and examination) of about 75 percent. That expected rate, coupled with the additional concern about unit non-response at various stages of the survey, fueled the need to conduct research aimed at addressing the potential problem of unit and item non-response.

49. As a consequence, beginning in 1992, a selected group of statisticians conducted a feasibility study to apply multiple imputation techniques to compensate for unit and item non-response in NHANES III. [14].

50. Multiple imputation, a state-of-the art methodology developed by Rubin [15], is a statistical technique in which missing data are replaced by several sets of plausible, alternative simulated values.

51. Using this methodology, each data set contains different simulated values for missing data. Each data set is analyzed separately, but in the same manner, and the analytic results are combined by simple rules to produce estimates, standard errors, and confidence intervals that incorporate appropriate levels of missing data uncertainty. For example, since body measurements in NHANES III are highly correlated, one missing measurement can be accurately predicted from other measurements with little variation across the multiple data sets. If, however, no body measurements exist, the variability across data sets will be much greater, producing measurement estimates similar to those that would be obtained from five random samples of the same demographic population [14].

52. In addition to comparing the relative merits of multiple imputation with those of single imputation, another key issue in considering the use of multiple imputation was determining the optimal number of imputations. As part of the feasibility study, this issue was investigated while comparing different imputation techniques [16,17,18].

53. Although NHANES III included hundreds of variables, the overall feasibility study, which was comprised of a number of objective specific research studies, was limited to a relatively small number of NHANES III variables in the imputation models. Although included variables came from the survey screener, the household interview, and the physical examination, the researchers cautioned that results could not be attributable to variables excluded from the models [14].

54. For example, the study by Ezzati-Rice et al [18] used selected demographic items and selected household interview variables to impute for six target examination variables: height, weight, diastolic blood pressure, systolic blood pressure, HDL cholesterol, and total cholesterol.

55. Using that test data set, the researchers investigated two single imputation methods using two closely related regression techniques, predictive mean matching and hot deck regression. A multiple imputation procedure was based on a multivariate model for mixed normal and categorical data, and used ten multiple imputations randomly generated by iterative simulations [16,19].

56. The data evaluated (from Phase I of NHANES III) under the three imputation methods produced similar point estimates and also preserved the marginal distribution of the variables and their relationships, thus showing no significant differences between any of the methods.

57. However, since Rubin [15] had earlier described the efficiency of multiple imputation in terms of its effect on variance estimates, the investigators found that, for most survey components, the increase in the variance using multiple imputation, as compared with single imputation, was reduced by a factor equal to the number of multiple imputations, with optimal reduction occurring when five multiple imputations were used.

XI. CONCLUSIONS

58. Data editing practices at NCHS are quite extensive. In documenting these practices, however, it is clear that certain data systems lend themselves to certain processing operations. The registration systems, for example, are set up to receive electronic submission of data. Thus, machine edits, using both hot deck and cold deck procedures for selected demographic variables, work well. The population based and provider based surveys, on the other hand, require more sophisticated procedures (e.g., sample case weight inflation, post-stratification adjustment, etc.) for imputation. Some additional major practices are summarized below.

- All NCHS data systems observe quality standards (e.g., error tolerances) in determining the level of data that can be disseminated.
- All NCHS data systems use one or a combination of micro, macro, or statistical editing.
- The majority of NCHS data systems identify imputed items, though fewer than half provide the percentage of a particular item that has been imputed.
- Virtually all NCHS data systems have an audit trail.
- Virtually all NCHS data systems monitor their automated editing procedures.
- More than two-thirds of NCHS data systems report item non-response rates under 5 percent.

59. With regard to the use of multiple imputation in NHANES III, the following observations were noted:

- Though close, multiple imputation may give a slightly better point estimate than single imputation due to sampling from the empirical distribution rather than assigning a probability mass to a single "donor." In particular, with small domains, single imputation could impute an outlying observation, whereas multiple imputation is less likely to introduce a bias, however slight, due to influential observations being imputed. However, this is not a significant factor if the number of missing items is large.

- Multiple imputation gives better estimates of sampling error (i.e., variance). Single imputations falsely reduce the variance because standard software packages recognize the imputed value as a legitimate value, while multiple imputations preserve the variability.

- Five multiple imputations are "empirically" optimal. Fewer and the variability is not sufficiently preserved; more imputations add little to the bias/variance comparisons.

- Multiple imputations are built on multivariate modeling strategy that preserves relationships in the data; single imputation on univariate data items can lead to illogical results in cross-comparison.

References

[1] Statistical Policy Working Paper 18: "Data Editing in Federal Statistical Agencies," Statistical Policy Office, Office of Information and Regulatory Affairs; Office of Management and Budget, Washington, DC, May 1990

[2] National Center for Health Statistics: "Data Editing at the National Center for Health Statistics, "Internal Report, Hyattsville, MD, November 1994

[3] National Center for Health Statistics: "Data Systems of the National Center for Health Statistics," Vital and Health Statistics, Series 1-No. 16, Hyattsville, MD, December 1981

[4] National Center for Health Statistics: Data Systems of the National Center for Health Statistics," Vital and Health Statistics, Series 1-No. 23, Hyattsville, MD, March 1989

[5] National Center for Health Statistics; "Origin, Program and Operation of the U.S. National Health Survey, "Vital and Health Statistics, Series 1 - No.1, Rockville, MD, August 1963.

[6] National Center for Health Statistics: "Cycle I of the Health Examination Survey, Sample and Response, United States, 1960-62, "Vital and Health Statistics, Series 11-No. 1, Rockville, MD, May 1965.

[7] National Center for Health Statistics: Sample Design: Third National Health and Nutrition

Examination Survey, " Vital and Health Statistics, Series 2-No. 113, Hyattsville, MD, September 1992.

[8] National Center for Health Statistics: "National Survey of Family Growth, Cycle I, "Vital and Health Statistics, Series 2-No. 76, Rockville, MD, June 1977.

[9] www.cdc.gov/nis

[10] Blumburg, S., Olson, L. Osbourn, L.,Srinath, K., and Harrison, H. "Design and Operation of National Survey of Early Childhood Health, 2000." Forthcoming.

[11] Grandquist, L. " A Report of the Main Features of a Macro-Editing Procedure which is used in Statistics Sweden for detecting errors in individual observations". Report presented at the Data Editing Joint Group, Statistics Sweden, 1987.

[12] Greenberg, B., and Surdi, R. "A Flexible and Interactive Edit and Imputation System for Ratio Edits". Proceedings of the Survey Research Methods Section of the American Statistical Association, Washington, D.C., 1984.

[13] National Center for Health Statistical, "A Statical Methodology for Analyzing Data from a Complex Survey, the first National Health and Nutrition Examination Survey, "Vital and Health Statistics, Series 2-No 92, Hyattsville, MD., September 1982.

[14] The National Center for Health Statistics, " The Third National Health and Nutrition Examination Survey, (NHANES III 1988-1994); Multiple Imputed Data Sets on CD-Rom Vital and Health Statistics , (Series 11, No. 7A, Hyattsville MD. June 2001.

[15] Rubin, D.B. Multiple Imputation for Non-Response in Surveys, New York, Wiley, 1987.

[16] Khare,M, Little R., Rubin, D, and Schafer, J "Multiple Imputation of NHANES III " Proceedings of the Survey Research Methods Section of the American Statistical Association, Washington D.C., 1993.

[17] Ezzati-Rice, T., Fahimi, M., Judkins, D, and Khare, M. "Serial Imputation of NHANES III with Mixed Regression and Hot-Deck Models," Proceedings of the Survey Research Methods Section of the American Statistical Association, Washington, D.C., 1993.

[18] Ezzati-Rice, T., Khare, M., Rubin, D., Little, R., and Schafer, J. " A Comparison of Imputation Techniques in the Third National Health and Nutrition Examination Survey." Proceedings the Survey Research Methods Section.

[19] Schafer, J., Khare, M., Ezzati-Rice, T. "Multiple Imputation of Missing Data in NHANES III." Proceedings of the 1993 Annual Research Conference, U.S. Bureau of Census, Washington, D.C.,1993.