WP.38
English only

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (iv): Data editing by respondents and data suppliers

Source Point Data Editing in Health Surveys

Supporting Paper

Submitted by the National Center for Health Statistics,USA[1]

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. (See Table 1)

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:

3. Procedures 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. Source point data editing, the subject of this paper, 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. This time

[1] Prepared by Kenneth Harris, Kwh1@cdc.gov.

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.

II. Source Point Data Editing Procedures

5. Most of the NCHS data systems implement the majority of their data editing procedures after data entry. In fact, NCHS's population based surveys are the only ones that, as a group, have instituted significant data editing procedures at or near the point of data collection. Just one provider based survey, the National Nursing Home Survey, and the lone Followup/Followback Survey, the Longitudinal Study of Aging II, also use source point data editing. These data systems use automated procedures during data collection which identify problems that offer the opportunity to recontact the data suppliers in a timely fashion. A total of eighteen specific automated software checks were identified as being used by the NCHS data systems. These automated checks are shown in Table 2.

6. This paper will describe the kinds of source point data editing procedures used by National Center for Health Statistics surveys.

III. Registration Systems

7. The vital event registration systems cover three vital events: mortality, fetal mortality, and natality. In addition, the Linked Birth and Infant Death Data Set is based on data obtained from the Mortality and Natality Registration Systems. For each of these systems, data are obtained from certificates and reports filed in state registration offices and registration offices of selected cities and other areas.

8. For each of its registration systems, NCHS monitors the quality of demographic and medical data on tapes received from the states 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. It should also be noted that the natality and mortality files, which are input to the linked file, have already passed rigorous quality control standards. Recontact with state registration offices is generally limited to cases where a cause of death is classified as a "rare" cause and NCHS must verify the accuracy of the classification.

9. Vital records and reports originate with private citizens—members of the families affected by the events, their physicians, funeral directors, and others. The responsibilities of these individuals are defined in States' laws. Birth registration is the direct responsibility of the hospital of birth or the attendant at the birth (generally a physician or midwife). In the absence of an attendant, the parents of the child are responsible for registering the birth. While procedures vary from hospital to hospital, usually the personal information is obtained from the mother; medical information may be obtained from the chart or from a worksheet filled out by the birth attendant.

10. Death registration is the direct responsibility of the funeral director or person acting as such. The funeral director obtains the data required, other than the cause of death, from the decedent's family or other informant. The attending physician provides the cause and manner of death. If no physician was in attendance or if the death was due to other than

natural causes, the medical examiner or coroner will investigate the death and provide the cause and manner.

11. Reporting requirements vary from State to State. In general, the completed birth certificate must be filed with the State or local registrar within ten days of the birth; death certificates must be filed within three to five days of the death.

12. There are 54 registration areas of the United States (50 States, District of Columbia, New York City, Puerto Rico and The Virgin Islands) that submit data to NCHS in electronic form through the Vital Statistics Cooperative Program (VSCP). Under the terms of the VSCP contracts with NCHS, the registration area must incorporate NCHS specifications into their own procedure so that the resultant data files meet the needs of both NCHS and the registration area.

13. All registration areas provide coded data to NCHS in electronic form (compact discs, diskettes, or PC to PC transmission). These files include all the births and deaths registered within their jurisdiction for each calendar year. Transmittals take place at regular intervals and contain all records received and initially processed in the State office since that last transmittal to NCHS, regardless of the month of the occurrence of the event. A record need not be "perfect" to qualify for transmittal. Each regular data transmittal should contain all replacement records processed to date incorporating updated information from any source. For purposes of full utilization and release of the data, States are expected to transmit the majority of records within six months of occurrence and a complete and final version of all records by June 30 of the following year.

14. Automation at the data source is a critical element of the system. Electronic birth and death certificates (completed by hospitals, funeral directors, and physicians) facilitate record filing, reduce processing redundancies, increase timeliness, and can improve data quality. Experimentation with electronic birth certificates (EBC) began in the early 1980s. Currently, EBCs are in use in one or more hospitals in forty-eight states and approximately ninety-two percent of all U.S. births are registered electronically. However, most states operate a dual (electronic and paper) registration system, in part because state laws have often not kept pace with technology. The recent revision of the model law is intended to assist states in addressing these issues. Electronic death registration (EDR) has not progressed as rapidly, primarily because the death registration process is more complex than birth registration and involves many more data providers.

IV. Population Based Surveys

15. Five of the Center's data systems are classified as population based surveys. They are the National Health Interview Survey, National Health and Nutrition Examination Survey, the National Immunization Survey, the National Survey of Family Growth and the State and Local Area Integrated Telephone Survey.

V. National Health Interview Survey (NHIS)

16. The NHIS is a continuing nationwide sample survey 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.

17. Interviewers enter data on lap top computers that have Computer Assisted Personal Interviewing (CAPI) features. These include edits that alert the interviewers to impossible or unlikely entries.)

18. The in-house data editing system is highly formalized and well documented. Both automated and manual edits are applied to raw data files, primarily for "critical " fields such as age and sex, family structure, as well as race/ethnicity edits which are critical to retaining households in screening segments related to oversampling of black and Hispanic households. Each of these edits is double-checked by two or more staff members. Actual keyed response fields are evaluated by staff programmers to determine if a reasonably sufficient number of fields have valid data for the record to be retained.

VI. National Health and Nutrition Examination Survey (NHANES)

19. The NHANES obtains nationally representative information on the health and nutrition 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.

20. A comprehensive, continuous and tightly integrated Quality Assurance (QA)/Quality Control (QC) program has been instituted for NHANES 99+. QC is one of the most important aspects of any study, as the integrity of the conclusions drawn by the study is in large part determined by the quality of the data collected. These are two basis components to insuring data integrity: quality assurance and quality control. QA consists of those activities that take place before data collection or in improving and refining data collection, while QC consists of those activities that take place during and after data collection. Manual development, training/retraining before and during the survey, certification of examiners and feedback are part of the QA process. Component completion rate, validation of household interviews, contractor and sub-contractor debriefings, examiner performance, reliability and validity, mobile examination center (MEC) examination flow and equipment performance are part of the QC process.

VII National Immunization Survey (NIS)

21. The NIS collects vaccination information on children 19 to 35 months of age living in the United States that allows coverage rates to be monitored at national, state, and local area levels. Additionally, the objectives of the NIS are to assist the CDC in allocating resources to states for the purpose of increasing coverage rates, to identify subpopulations and/or geographic areas in which rates are low, and to provide a data base for epidemiological research.

22. The NIS collects immunization data from two sources-a telephone survey of households and a mail survey immunization providers identified by household respondents. For the household survey, the NIS employs a list-assisted random-digit-dialing (RDD) sample design.

23. After data collection ends, the NIS combines the data collected from household respondents and from vaccination providers in a comprehensive analytical file. This file is prepared to be as accurate as possible.

24. Even though the CATI (Computer Assisted Telephone Interviewing) system makes numerous checks during data collection, a final editing process identifies any remaining data inconsistencies and takes steps to reduce or eliminate them. Once the CATI production files are passed to the data preparation stage, various household-and child-level files are produced by extracting specific fields from CATI data.

25. A master look-up database; constructed from the questionnaire, contains information about each field, including allowable ranges of responses. The master database is maintained, reviewed, and updated each quarter. At the end of each quarter of data collection the raw data are matched against this master database, and a report is then reviewed by a senior project analyst for resolution.

26. During the data collection phase of the Provider Record Check Study, no attempt is made to identify and eliminate duplicate records. Duplicate records often result when a reminder call prompts a mail or fax return of a form that is already in the mail. All forms received are checked in a sent for data entry. After entry the data file is unduplicated to remove such records. The process is structured in a way that maintains the integrity of the child-provider pairs and ensures that all pertinent information for each child is keep intact.

27. Each file is subjected to a detailed check for out-of-range values and sources of missing data, to reduce the potential for error in the data entry process. Out-of-range data often indicate incorrect vaccination dates. Problem cases are pulled and reviewed manually by a senior analyst. Problems are resolved using all available data sources: household information, provider data, partnering agency experts, and recommended vaccination schedule. As in the NIS household file editing process, the analyst helps to identify problems or trends in the data that can lead to the reduction of error.

28. Any problems that cannot be resolved remain in the file until the household data and the provider data are combined.

VIII. National Survey of Family Growth (NSFG)

29. The NSFG is a periodic nationally representative household survey of women of reproductive age (15-44 years). The survey collects data on fertility and infertility, family planning, and related aspects of maternal and infant health. In the most recent NSFG (2002) men were included for the first time.

30. CAPI Edits.—In1995 and 2002, interviewing was conducted by Computer-Assisted Personal Interviewing, or CAPI, in which the interviewer read the questions from a laplop computer, and entered the respondent's answer into the computer. This process made it possible to improve the quality of the data by designing edits in advance of data collection.

31. Post-interview editing – These edit checks are themselves tested during the data editing process. After the data are received from the contractor, some of these edit checks are performed again to ensure that they operated properly.

32. NSFG and contracting staff examine the data to identify instances where a "not ascertained" code is necessary, due to a deviation from the routing specified in the program (for example, due to a suppression of an edit check's "error message").

IX. State and Local Area Integrated Telephone Survey (SLAITS)

33. The State and Local Area Integrated Telephone Survey (SLAITS) was designed as a household telephone survey interview on a varying set of health and welfare related topics. The content and the sample design for each SLAITS module are customized to meet the needs of survey sponsors, and generally are designed to describe the health, health care, or welfare of a particular population at one point in time.

34. SLAITS is a random digital dial survey (RDD). The telephone numbers generated for SLAITS are tested electronically for being in scope and working numbers.

35. The prepared sample of telephone numbers is checked to ensure that it meets the sample design specifications. The sample is monitored on a daily basis to ensure that the pace of data collection is consistent across the data collection period and to prevent the unnecessary release of excess cases. Daily analyses of the dynamics in the sample are produced to assist in timely sample management decision-making.

36. Data Quality Control: The CATI system is programmed to help ensure complete and accurate data collection, using automated data checks techniques, such as response-value range checks and consistency edits, during the interview process. These features enable interviewers to obtain needed clarifications while still on the telephone with the respondent. Throughout data collection, interview data are reviewed for consistency between fields, appropriate response-value ranges, skip logic patterns, and missing information.

37. Concurrent with the development of the CATI questionnaire for the data collection phase, a detailed plan for checking and editing the data in the CATI instrument is developed. The intention is to design into the CATI software consistency checks across data elements, valid range codes, and a method to identify incorrect codes entered by interviewers. To the extent that the CATI software could be developed to perform these tasks, the efficiency of post survey data cleaning and processing is increased.

38. The CATI system is designed to perform a number of edits as an interviewer enters data into the computer system. These edits deal with errors that could be reconciled while the respondent is on the telephone and focused, in particular, on items critical to the conduct of the study. The CATI edit specifications are designed to correct respondent error during the interview (for example, a respondent saying two children under three years of age lived in the household, but only listing one name on the roster) and to identify and correct data-entry error by interviewers (for example, a 9-month old child is reported as being introduced to solid foods when she was 4-months-old, but the interviewer attempts to enter 14 months). To the extent possible without making the CATI system overly complicated, out-of-range and inconsistent responses result in a warning screen for the benefit of the interviewers, who are trained to correct errors as they occur. These messages are designed primarily to prevent data entry errors and respondent errors and not to change respondents who give logically inconsistent responses.

X. Provider Based Surveys

39. Five NCHS data systems form the family of provider based surveys, collectively called the National Health Care Survey (NCHS). Included here are the National Hospital Discharge Survey, National Ambulatory Medical Care Survey, National Hospital Ambulatory Medical Care Survey, National Nursing Home Survey, and the National Home and Hospice Care Survey. 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. The National Nursing Home Survey currently is the only one that uses source point data editing procedures.

XI. National Nursing Home Survey (NNHS)

40. The NNHS is a national probability sample survey of nursing homes, their residents and staffs. Data about facilities include basic characteristics such as size, ownership, Medicare/Medicaid certification, staff, occupancy rate, days of care provided and financial characteristics. Data about the patient, both current and discharged residents, include: basic demographics, marital status, place of residence prior to admission, health status, services received, and for discharges, the outcome of care.

41. Interviewers enter data on lap top computers that have CAPI features, including edits that alert the interviewer to impossible or unlikely entries.

XII. Followup/Followback Survey

42. Currently only one NCHS data system included in this report is classified as a Followup/Followback survey.

XIII. Longitudinal Study of Aging II (LSOA)

43. The LSOA is a family of surveys based on the Supplement of Aging (SOA) to the 1984 National Health Interview Survey. The SOA was designed to obtain extensive information on family structure and frequency of contacts with children; housing (including barriers to movement, length of time in residence, ownership, and rental information); use of community and social supports; occupation and retirement (including sources of retirement income); ability to perform work-related functions; conditions and impairments; functional limitations (activities of daily living and instrumental activities of daily living) and providers of help in those activities.

44. Interviewers enter data on lap top computers that have CAPI features, including edits that alert the interviewer to impossible or unlikely entries.

XIV. Conclusion

45. Source point data editing is a valuable tool for improving the completeness and quality of data colleted in NCHS health surveys. New procedures are regularly being tested and evaluated. Because of certain data collection constraints with regard to some provider based surveys, it is likely that NCHS will never have 100 percent coverage. Nevertheless, the number of surveys employing source point data editing is expected to increase over time.

References

Statistical Policy Working Paper 18: “Data Editing on Federal Statistical agencies,” Statistical Policy Officer, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, D.C. May 1990.

Table 1 NCHS Health Surveys

Vital Event Registration Systems (4)
Mortality
Fetal Mortality
Natality
Linked Birth and Infant Death

Population Based Surveys (5)
National Health Interview Survey (NHIS)
National Health and Nutrition Examination Survey (NHANES)
National Immunization Survey (NIS)
National Survey of Family Growth (NSFG)
State and Local are integrated Telephone Survey (SLAITS)

Provider Based Surveys (5)
National Hospital Discharge Survey (NHDS)
National Ambulatory and Medical Care Survey (NAMCS)
National Hospital Ambulatory and Medical Care Survey (NHAMCS)
National Home and Hospice Care Survey (NHHCS)
National Nursing Home Survey (NNHS)

Followup/Followback Survey (1)
Longitudinal Study of Aging II (LSOA)

Table 2 Frequency of Selected Source Point Data Editing Practices Among NCHS Data Systems

es your software provide the lowing general features?	**Mortality Reg ***	**NHIS**	**NHANES**	**NIS**	**NSFG**	**SLAITS**	**NNHS**	**LSOA II**
tomated Skips	NA	Y	Y	Y	Y	Y	Y	Y
cking up and changing data ring the interview	NA	Y	Y	Y	Y	Y	Y	Y
tomatic adjusting of skip tterns for changed entries	NA	Y	Y	Y	Y	Y	Y	Y
ility to access and change data after interview is over but ring SPDE timeframe	NA	N	Y	Y	Y	Y	Y	Y
-entering and correcting ormation about existing mbers in a roster	Y	Y	Y	Y	Y	Y	Y	Y
-entering and adding members a roster	Y	Y	Y	N	Y	N	Y	Y
-entering and deleting members m a roster	Y	Y	Y	N	N	N	Y	Y
ntext sensitive help	N	Y	Y	Y	Y	Y	Y	Y
rcing correction before ceeding if a data entry is valid	Y	Y	Y	Y	Y	Y	Y	Y
arning of unusual data entries	Y	Y	N	Y	Y	Y	Y	N
rced confirmation before storing ta	N	N	N	N	?	N	Y	Y
lowing the interviewer to enter mments on each question	NA	Y	N	N	Y	N	Y	Y
lowing other comments and tes	NA	Y	Y	Y	N	Y	Y	Y
t of range values	Y	Y	Y	Y	Y	Y	Y	Y
consistency	Y	Y	Y	Y	Y	Y	Y	Y
gal blanks	N	N	Y	N	Y	N	Y	Y
valid dates	Y	Y	Y	Y	Y	Y	Y	Y
valid characters (e.g. alpha aracters in a field that requires a meric answer)	Y	Y	Y	Y	Y	Y	Y	Y
e respondents even recontacted er data collection is over?	Y	Y	Y	Y	Y	Y	N	Y
es your survey recontact pondents to check for data sified by the interviewer?	NA	Y	Y	Y	Y	Y	N	N

*Also applies to Fetal Mortality and Natality