SAGE Secondary Data Analysis

Obtaining and Evaluating Data Sets for Secondary Analysis in Nursing Research

Contributors: Ann F. Jacobson & Patti Hamilton & James Galloway Western
Editors: John Goodwin
Book Title: SAGE Secondary Data Analysis
Chapter Title: "Obtaining and Evaluating Data Sets for Secondary Analysis in Nursing Research"
Pub. Date: August 1993
Access Date: October 17, 2013
Publishing Company: SAGE Publications Ltd
City: London
Print ISBN: 9781446246900
Online ISBN: 9781446268544
DOI: http://dx.doi.org/10.4135/9781446268544

http://dx.doi.org/10.1177/01934599301500407http://
dx.doi.org/10.1177/019394599301500407
[p. v4-29 ↓]

Obtaining and Evaluating Data Sets for Secondary Analysis in Nursing Research

http://dx.doi.org/10.1177/019394599301500407
Encoding from PDF of original work

‘Obtaining and Evaluating Data Sets for Secondary Analysis in Nursing Research’, Ann F. Jacobson Patti Hamilton James Galloway Western *Journal of Nursing Research,* vol. 15 no. (4) (1993): pp. 483–494. Published by SAGE Publications, Inc. Reprinted with permission.

Recent articles in nursing journals have advanced cogent arguments for using secondary analysis of existing data as a method of research (Aaronson, 1990; McArt & McDougal, 1985). As standards for quality in research grow more stringent and competition for scarce funds intensifies, the economy of secondary data analysis compared to traditional procedures is becoming readily apparent. For as little as $100 (U.S.), a researcher can obtain data on hundreds of health-related variables from large cross-national samples, such as those from the National Medical Expenditure Survey (Edwards & Edwards, 1989) and the Older Americans Resources and Services (Duke University Center for the Study of Aging, 1979).

In secondary analysis, the researcher tests new hypotheses by using raw data that have been collected by someone else. Secondary analysis should not be confused with meta-analysis, in which the results from a number of studies in the field are statistically combined (Burns & Grove, 1987, p. 277).

Advantages and Limitations of Secondary Analysis

Advantages of secondary analysis, compared to primary data collection, have been well documented (Aaronson, 1990; Bowering, 1984; David, 1991; Jacob, 1984; Kiecolt & Nathan, 1985; Miller, 1982; Sieber, 1991). Few individual investigators have the resources to draw a random sample or to include more **[p. v4-30 ↓]** than several hundred participants, whereas large national surveys may have a randomly selected sample of 1,000 or more. Large random samples are frequently desired in research to increase statistical power and generalizability of findings, meet the assumptions of many statistical tests (especially multivariate ones), and provide narrow confidence intervals for interpretation. (Large samples can, however, generate statistical significance when substantive or clinical significance is dubious.) The use of paid, trained interviewers in large-scale studies can increase reliability and validity *(Conference Proceedings: Health Survey Research Methods*, 1989).

A major limitation of secondary data analysis is lack of control over how a data set was conceived, generated, or recorded. Lack of involvement during the data collection phase limits the researcher's insight into hidden factors that may have influenced the study's outcome. Consequently, the secondary analyst is at greater risk for drawing invalid conclusions that arise from misinterpretation of findings.

Locating Sources of Data for Secondary Analysis

Despite the appeal of secondary data analysis, lack of knowledge of available data sets has prevented nursing from using the technique as frequently as other disciplines (Aaronson, 1990). Information scientists have identified the major problems with user access to data sets but admit that solutions have been difficult to implement (Wenzel, 1988).

Usual sources of precollected data include individual researchers, university archives or systems, government repositories, foreign or international agencies, and independent archives. Members of the first group, individual researchers, are usually discovered via local informal networks. Nurse researchers, for example, often make their data available for students' or colleagues' research projects. Although sources in the remaining four categories have broader accessibility, they may be more difficult to locate.

To develop a list of accessible sources of data files relevant to nurse researchers, the authors, using keywords "secondary data analysis" or "data archives," searched the following indexes from 1980 to 1991 through the DIALOG Information Services: Biosis Previews, ERIC, Federal Research in Progress, Health Periodicals Database, Medline, Nursing and Allied Health (CINAHL), and PsycINFO. In addition, a reference well-known to information specialists, the *1990 Encyclopedia of Information Systems and Services* (Lucas, 1990), was consulted. This reference is a compendium of more than 4,400 organizations involved in the production and distribution of electronic data. From these, a list of sources of data with potential interest to nurses was developed.

The organizations were contacted by telephone between January 1991 and July 1992 for information about their holdings and their availability. Of the 23 organizations contacted, 17 had few or no access restrictions to **[p. v4-31 ↓]** data collected on variables of interest to nurse researchers. The names of the organizations and a summary of their holdings are presented in Table 1.

Although this list is extensive, it should not be considered exhaustive. Until a more systematic and comprehensive method for identifying suitable data bases for nursing is established, the technique will remain imperfect.

Table 1: Selected data sources

Source	Holdings
University Archives	
Duke University Medical Center Center for the Study of Aging and Human Development Survey Data Laboratory Box 3003 Durham, NC (919) 684-3204	Machine-readable data from federal, private, and university studies, including all aspects of aging and adult development: health, economics, mental health, life satisfaction, functional status, and others.
Mississippi State University Social Science Research Center Data Archives P.O. Box 5387 Mississippi State, MS 39762 (601) 325-7127	Data from the Social Science Research Center on a range of topics, including drug and substance abuse.
University of Alberta Computing Services Data Library 4-15 Cameron Edmonton, AB, Canada T6G 2H1	Machine-readable data files from more than 700 studies, consisting of surveys, public opinion polls, social and political science data, and clinical trials, dating from 1977 to the present; special collection of quality of life research; list of other Canadian data sources available on request.
University of Alberta Population Research Laboratory Edmonton, AB, Canada T6G 2H4 (403) 492-4659	Census of Canada summary tapes and a number of surveys done in Alberta and the Edmonton area; the laboratory provides retrieval services from the Census of Canada tapes and will also supply tape copies of its data sets with the permission of the director.
University of Connecticut Institute for Social Inquiry Roper Center for Public Opinion Research User Services Development 341 Mansfield Road, Room 421 Box U-164 Storrs, CT 06268, (203) 486-4440	Machine-readable data files on social science-related studies, including numerous health-related studies; Gallup and Roper report data from the 1930s to the present.
University of Detroit Computerized Folklore Archive 4001 W. McNichols Road Detroit, MI 48221 (313) 927-1081	36,000 items of folklore collected by students in the University's folklore classes; items such as legends, folk tales, folk songs, folk medicine, and superstitions are stored on cards, tape, and disk and indexed by subject, group, time, place, and other headings.
University of New Hampshire State and Regional Indicators Archive 128 Horton Social Science Center Durham, NH 03824 (603) 862-1888	Data collected from samples in all 50 states for approximately 13,000 variables; catalog and file documentation available.
University of North Carolina Institute for Research in Social Science Social Science Data Library Manning Hall, Room 10 Chapel Hill, NC 27514 (919) 966-3346	More than 2,000 machine-readable data files representing 32 major categories in the areas of political science, social science, and economics; data acquired from Louis Harris public opinion surveys, government agencies, academic institutions, and individual contributors.

[p. v4-32 ↓]

Source	Holdings
University of Waterloo Department of Recreation Leisure Studies Data Bank Burt Matthews Hall 2113 200 University Avenue Waterloo, ON, Canada N2L 3G1 (519) 885-1211	More than 100 machine-readable data files created from data contributed by researchers worldwide; topics include sport, fitness, recreation, leisure, and related topics; catalog of holdings with study information available.
U. S. Government repositories	
National Technical Information Service 5258 Port Royal Road Springfield, VA 22161 (703) 487-4650	Data files and documentation from selected projects funded by the National Center for Health Services Research, Department of Health and Human Services; catalog and file documentation available.
Bureau of the Census Data Users Service Division Customer Services Branch Washington, DC 20233 (202) 763-4100	U. S. Census data.
International Source	
World Health Organization Division of Epidemiological Surveillance and Health Situation and Trend Assessment World Health Statistics Data Base 20 Avenue Appia CH-1211 Geneva 27, Switzerland Telephone 022 912111	Machine-readable data files containing health and demographic data from participating governments.
Independent Archives	
Institute for Aerobics Research Aerobics Center Longitudinal Study 12330 Preston Road Dallas, TX 75230 (214) 701-8001	Data base of longitudinal data collected on health behaviors, nutrition, and exercise and clinical assessment.
Institute for Child Behavior Research Data Bank 4182 Adams Avenue San Diego, CA 92116 (619) 281-7165	Machine-readable data files on more than 10,000 children with severe learning and behavior disorders, especially autism, throughout the world.
Inter-University Consortium for Political and Social Research Resource Development Institute for Social Research P.O. Box 1248 Ann Arbor, MI 48106 (313) 764-5199	More than 25,000 machine-readable files representing more than 6 million variables, in approximately 1,600 studies in political and social sciences; also houses the National Archive of Computerized Data on Aging; data files are free or at minimal cost to members, whereas nonmembers' cost to purchase files is higher; catalog and file documentation available.
RAND Corporation Computation Center Data Facility 1700 Main Street Santa Monica, CA 90406 (213) 393-0411	More than 500 machine-readable data bases, including approximately 43,000 files on a broad range of topics, including health; data are generated by RAND researchers or acquired from other agencies; file documentation available.
Sociometrics Corporation Data Archive on Adolescent Pregnancy and Pregnancy Prevention 170 State Street, Ste. 260 Los Altos, CA 94022 (415) 949-3282	Data from more than 110 studies available on tape or diskette from studies of sexual behavior of teenagers; topics include contraception, pregnancy, adolescent family life, and teen parenting; SPSS program statements and documentation accompany each file; catalog available.

[p. v4-33 ↓]

Aaronson (1990) recognized this difficulty and proposed the establishment of a national data repository for nursing as a solution.

In addition to the organizations listed in Table 1, local sources may hold data suitable for secondary analysis. For example, most major state and provincial universities maintain data files that may be available to researchers outside their system. State and local governments also can provide data files for users with more regionally defined interests. For example, Moon (1992) used state health department records to analyze the effect of certain demographic variables on infant birth weight in Oklahoma.

Evaluating Potential Data Sets

The investigator who hastily acquires a data set, based on a superficial understanding of the file's contents, may be frustrated and disappointed with unsuitable data. Therefore, the time spent on learning as much as possible about the data set and the study that produced it can save hours once the analysis is under way.

In researching a particular data set, an important caveat is that things might not be, and probably are not, as they first appear. Population characteristics, sampling methods, and instruments used are a few areas that may initially seem appropriate but, on further examination, be problematic for the secondary study. The following criteria are essential for evaluating a potential data set (others may apply, depending on the investigator's skills, resources, and research aims).

Technical Factors

Most data sets are made available in machine-readable data tape format that requires loading onto a mainframe or minicomputer prior to data analysis. The tape format must be compatible with certain characteristics of the local computer environment. These include magnetic recording codes (e.g., EBCDIC) and tape density (number of bytes per inch). Many data providers preformat tapes for use with one of the statistical packages (SPSS-X or SAS). Secondary data files frequently exceed the amounts of computer memory allotted to individuals in academic institutions; therefore, the researcher should anticipate the need for acquiring larger working and storage file spaces within the local computer environment.

Study Variables

A common pitfall in secondary analysis is assuming that the names of a data set's variables adequately reflect the secondary analyst's concepts of interest. Scrutiny of the conceptual and operational definitions of the study variables **[p. v4-34 ↓]** is essential in determining the appropriateness of a data set for the secondary analyst's purpose.

For example, health status of a study's participants is frequently of interest to a nurse researcher. Large surveys frequently use illness inventories, symptom checklists, or number of physician office visits as measures of health. The researcher must judge whether such data adequately fit the conceptual definition of health proposed for the secondary analysis.

Fortunately, problems arising from inadequate measures for conceptual variables are not insurmountable. Composites of items within the data set can be constructed as indicators of the secondary analyst's conceptual variable (Kiecolt & Nathan, 1985). Alternatively, raw data from several studies on random samples of the sample population can be combined to yield the desired set of variables (Jacob, 1984). Fortune and McBee (1984) have outlined specific techniques for variable construction in data file preparation.

Data Collection Procedures

As the secondary analyst was absent during the generation of data, meticulous examination of the study design and data collection methods is necessary. For example, Duke University Archive describes the OARS data set as representing a random sample of Medicare-eligible subjects in Cleveland, Ohio. However, additional documentation refers to two subsamples included in the study: Medicare-eligible participants and supplementary security income (SSI) recipients (U.S. General Accounting Office, n.d.). These two samples were originally defined due to a misconception that they were mutually exclusive (W. Laurie, personal communication, 1989), when in fact all SSI recipients are also Medicare eligible.

Elements of the study design and data collection methods used by the primary researchers that should be examined include the following:

Sufficient information to evaluate all of these criteria pertaining to data collection is seldom available, nor are all the criteria of equal importance. **[p. v4-35 ↓]** The investigator should consider the purpose of the data reanalysis as the ultimate guide in applying selected criteria for an individual study (for a detailed discussion of evaluating a data set's technical aspects, see David, 1991).

Documentation of Data Set

Documentation that accompanies the data set should include information about the organization or arrangement of the data tape fields and elements as well as the procedures used for coding missing data and formulas for weighting variables. Frequency tables of the original items should be provided in the documentation to allow the secondary analyst to assess accuracy of the data tape by comparison. Documentation should also include copies of the instruments used as well as a detailed description of the method. A complete description of data editing and coding procedures, along with error rates, permits the user to further evaluate data quality.

Age of the Data

The time at which the original data were collected is an important consideration not only in terms of determining currency of the data but also in assessing historical factors in operation during data collection. For example, studies conducted in the United States prior to 1983, pertaining to hospitalization-related variables, such as length of stay, would reflect pre-diagnosis-related group (DRG) phenomena. Unfortunately, not all historical factors are as obvious as this example. Substantive knowledge of the subject matter is the researcher's best insurance against threats to validity related to historical factors.

Availability of Contact Person

During a secondary analysis, questions arise that even the most extensive and complete documentation might not address. For example, the secondary analyst may discover from literature review that responses of a particular population are sensitive to the age or gender of the interviewer, but the documentation does not address an interviewer-subject assignment strategy. Availability of persons who were involved in the data collection and entry phase of the original study can be valuable to the secondary researcher with such concerns.

Assistance with Secondary Analysis

Successful completion of a secondary analysis project may require locating appropriate human resources and information systems. Often, research or reference librarians in university or large public libraries can assist researchers in the identification of organizations and libraries that offer data tapes and **[p. v4-36 ↓]** associated services, have print and online bibliographic listings, and provide guidance with technical aspects of secondary analysis.

Table 2: Organizations with an interest in secondary data analysis

Organization	*Membership information*
Association of Public Data Users (APDU) Princeton University Computing Center 87 Prospect Avenue Princeton, NJ 08544-1002 (609) 258-6025/6052	APDU members are users, producers, and distributors of federal, state, and local government statistical data, including government agencies, academic institutions, commercial vendors, nonprofit organizations, and individuals; publishes a monthly newsletter and holds an annual convention in Washington, DC; membership fee: individual, $125 yearly, and organization, $250 yearly.
International Association for Social Science Information Service and Technology (IASSIST) 1700 Main Street Santa Monica, CA 90406 (213) 393-0411	Members represent an international cooperative of data librarians, individuals, and groups using archived data; publishes the *IASSIST Quarterly*; membership fee: $20 yearly.
National Association of Health Data Organizations (NAHDO) 254B N. Washington Street Falls Church, VA 22046 (703) 532-3282	NAHDO is a membership organization with members from the public and private sector; the Association's stated goal is to promote standardization and uniformity of health data, and access to health data; publishes *NAHDO NEWS*, holds annual meetings, and sponsors seminars; membership fee: associate, $5,000 yearly, supporting, $2,500 yearly, public, $2,000 yearly, and individual, $125 yearly.

Many of the data sources listed in Table 1 provide technical and statistical assistance and may even furnish lists of previous users of certain data sets. In addition, the organizations named in Table 2 are open to members interested in issues surrounding the use of existing data sets.

Another important resource for secondary analysts are publications that address methodological and/or statistical considerations for users of published and/or survey data. Many publications cited in the References Section discuss strategies and issues helpful to the secondary analyst.

Examples of Secondary Analysis in Nursing Research

Aaronson (1990) predicted an increase of secondary data analysis in nursing research and cited several recent examples of this trend (Aaronson, 1989; Loveland-Cherry, Youngblut, & Leidy, 1989; Yarcheski & Mahon, 1989). To locate other examples, an online Medline search dating from 1986 to the present was conducted in July 1992 for citations with the words “secondary data analys?” (where “?” denotes truncation of the root “analys”) in a title or abstract. The search identified 11 citations, 3 of which appeared in nursing journals (Gleit & Graham, 1989; Herron, 1989; Palmer, German, & Ouslander, 1991). (Aaronson's examples were not revealed by the Medline search because *Nursing Research* does not provide abstracts to bibliographic data bases and “secondary data analysis” did not appear in the title of any of these examples.)

[p. v4-37 ↓]

Summary

Secondary analysis of existing data offers many advantages to the nurse researcher. Data from large-scale studies may be reanalyzed and refined by secondary analysts with a fresh perspective, thus enhancing the original study's contribution to scientific knowledge. High-quality data can be obtained for comparatively little expenditure of time and money. The secondary analyst, however, must exercise care in evaluating and analyzing a data set to maximize the internal and external validity of the reanalysis.

Because the secondary analyst's lack of involvement in data collection procedures may decrease insight into the original study's limitations, vigilant skepticism should accompany all phases of the research process in secondary data analysis, just as it should in all other research. Miller (1982) advised, “Begin by assuming the worst and seek out the same kinds of information about sample selection procedures, sample size, response rates, field procedures, and coding conventions that you would insist on if you were collecting your own data” (p. 722).

By systematically evaluating potential data sets according to rigorous predetermined criteria, the nurse researcher can minimize the possible pitfalls inherent in secondary analysis. On the other hand, investigators who use secondary sources appropriately can make significant contributions to nursing science at less cost than that engendered by traditional research methods.

References

Aaronson L. S. Perceived and received support: Effects on health behavior during pregnancy. Nursing Research, (1989). vol. 38, pp. 4–9.

Aaronson L. S. Needed: A national repository of nursing research data (Commentary). Nursing Research, (1990). vol. 39, pp. 311–313.

Bowering, D. J. (ed.). (1984). Secondary analysis of available data bases. San Francisco: Jossey-Bass.

Burns, N., & Grove, S. K. (1987). The practice of nursing research: Conduct, critique, and utilization. Philadelphia: Saunders.

Conference proceedings: Health survey research methods (DHHS Publication No. PHS 89–3447). (1989). Rockville, MD: National Center for Health Services Research.

David, M. (1991). The science of data sharing: Documentation. In J. E. Sieber (ed.), Sharing social science data. Newbury Park, CA: Sage.

Duke University Center for the Study of Aging. (1979). Multidimensional functional assessment: The OARS methodology. Durham, NC: Author.

Edwards, W., & Edwards, B. (1989). Questionnaires and data collection methods for the institutional population component (DHHS Publication No. PHS 89–3440). National Medical Expenditure Survey Methods 1, National Center for Health Services Research and Health Care Technology Assessment. Rockville, MD: Public Health Service.

Fortune J. C. and McBee J. K. Considerations and methodology for the preparation of data files. New Directions for Program Evaluation, (1984). vol. 22, pp. 27–49.

Gleit C. and Graham B. Secondary data analysis: a valuable resource. Nursing Research, (1989). vol. 38, pp. 380–381.

Herron D. G. Secondary data analysis: Research method for the clinical nurse specialist. Clinical Nurse Specialist, (1989). vol. 3 (2), pp. 66–69.

Jacob, H. (1984). Using published data: Errors and remedies. Beverly Hills, CA: Sage.

Kiecolt, K. J., & Nathan, L. E. (1985). Secondary analysis of survey data. Beverly Hills, CA: Sage.

Loveland-Cherry C. J. , Youngblut J. M. , and Leidy N. K. A psychometric analysis of the family environment scale. Nursing Research, (1989). vol. 38, pp. 262–268.

Lucas, A. (ed.). (1990). 1990 encyclopedia of information systems and services. Detroit: Gale Research.

McArt E. and McDougal L. Secondary data analysis — A new approach to nursing research. Image, (1985). vol. 17, pp. 54–57.

Miller J. D. Secondary analysis and science education research. Journal of Research in Science and Teaching, (1982). vol. 19, pp. 719–725.

Moon, V. M. (1992). Models of low birthweight. Unpublished manuscript, Texas Woman's University, College of Nursing, Denton.

Palmer M. H. , German P. S. , and Ouslander J. G. Risk factors for urinary incontinence one year after nursing home admission. Research in Nursing and Health, (1991). vol. 14, pp. 405–412.

Sieber, J. E. (1991). Introduction: Sharing social science data. In J. E. Sieber (ed.), Sharing social science data. Newbury Park, CA: Sage.

U.S. General Accounting Office. (n.d.). Study of the well-being of older people in Cleveland, Ohio. Survey file codebook: Volume 1. (Available from U.S. General Accounting Office, Room 2933, 1240 East Ninth Street, Cleveland, OH 44199)

Wenzel P. Microcomputer-based access to machine-readable numeric databases. Reference Services Review, (1988). vol. 16, pp. 51–55.

Yarcheski A. and Mahon N. E. A causal model of positive health practices: The relationship between approach and replication. Nursing Research, (1989). vol. 38, pp. 88–93.

http://dx.doi.org/10.1177/019394599301500407