2007

Issues associated with secondary analysis of population health data

Sandra C. Garmon Bibb
Uniformed Services University of the Health Sciences

ELSEVIER

Applied Nursing Research 20 (2007) 94–99

Applied Nursing Research

www.elsevier.com/locate/apnr

Clinical Methods

Issues associated with secondary analysis of population health data

Sandra C. Garmon Bibb, DNSc, RN*

Department of Health Systems, Risk, and Contingency Management, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA
Graduate School of Nursing, Uniformed Services University of the Health Sciences, Silver Spring, MD 20904, USA

Received 23 May 2005; revised 31 July 2005; accepted 18 February 2006

Abstract Analysis of population health data is crucial for the success of population health. Secondary analysis of data contained in publicly available clinical, administrative, and national health survey databases is a cost-effective and scientific method for generating data to support population health. There are several issues associated with secondary analysis of population health data that must be addressed if analyses are to be expedient, economical, and reliable. This article outlines the value of using secondary analysis to generate population health data, discusses issues associated with secondary analysis of population health data, and offers suggestions for addressing these issues.

1. Introduction

The study and analysis of population health data are crucial to the success of population health. Population health emphasizes assessment of health status, acknowledgement of multiple determinants of health, use of evidence-based interventions, and evaluation of health outcomes (Fos & Fine, 2005; Kindig & Stoddart, 2003). A population is defined in terms of the specific group under study and can be divided into categories based on demographic characteristics, health status, health care needs, and disease patterns and trends. Population health data provide descriptive information relating to these categories and are useful in planning, designing, operating, and evaluating health services. Although population health data can be generated in many ways, secondary analysis of data contained in publicly available clinical, administrative, and national health survey databases is an efficient and expedient mechanism for producing data to support population health.

Secondary analysis is the reanalysis of existing data or analysis of data collected for reasons other than research (Vogt, 2005). To date, use of secondary analysis as a research methodology in support of population health has been undertaken primarily by health service researchers. Nursing is just beginning to contribute to the body of health care research by means of health services research (Castle, 2003; Smaldone & Connor, 2003). Health services research is a multidisciplinary field of scientific inquiry that is concerned with the identification of health care needs and the provision, effectiveness, and use of health services (Bowling, 2002; Hughes, 2004; Jones & Lusk, 2002). The goals of population health and the foci of health services research are interrelated, and during the last 20 years, increased emphasis has been placed on population health in the United States. The U.S. Department of Health and Human Services' (HHS) *Healthy People* documents, the Institute of Medicine's *Crossing the Quality Chasm* publications, and the Agency for Healthcare Research and Quality's (AHRQ) *National Healthcare Quality* reports highlight the importance of assessment of population health and implementation of evidence-based interventions. These organizations rely heavily on secondary analysis of data contained in administrative, clinical, and national health survey databases (Agency for Healthcare Research and Quality [AHRQ], 2004; U.S. Department of Health and Services, 2000). This reliance is consistent with a growing national trend toward expedient, economical research and increased emphases on health services research (Bowling, 2002; Hughes, 2004; Jones & Lusk, 2002). Secondary analysis of population health data is one way to increase

* Graduate School of Nursing, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA. Tel.: +1 301 625 8896 (home), +1 301 295 1206 (work); fax: +1 301 625 8623 (home), +1 301 295 1707 (work).
E-mail address: sbibb@usuhs.mil.

0897-1897/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.apnr.2006.02.003

nursing's involvement with health services research while generating knowledge necessary for improving the health of populations and meeting current health care demands.

Secondary analysis is a practical, cost-effective, and scientific method and offers ready access to large data sets with multiple variables. However, secondary analysis of population health data is associated with several issues, which, if not addressed, can impact the expediency, economy, and reliability of data analysis. The purpose of this article is to outline the value of using secondary analysis to generate population health data, discuss issues associated with secondary analysis of population health data, and offer suggestions for addressing these issues.

2. Secondary analysis of population health data

The term *population health* is fairly new, but the population health approach and emphasis on the use of administrative, clinical, and national health survey data sets to generate population health data are not. *Healthy People 2010* outlines two broad goals for the nation's health: increasing quality and years of life and eliminating health disparities (HHS, 2000). This document identifies biology, behaviors, social environment, physical environment, policies and interventions, and access to quality health care as determinants of health and theorizes that health is influenced by these determinants either individually or through their interactions with each other. *Healthy People 2010* emphasizes the importance of assessing, measuring, and monitoring population health and identifies a variety of sources of existing data that can be used to assess health and measure outcomes. These data sources include several national health surveys such as the National Health and Nutrition Examination Survey, several clinical databases such as the United States Renal Data System, and several administrative data systems such as the National Vital Statistics System—Mortality (HHS, 2005).

Crossing the Quality Chasm: A New Health System for the 21st Century stresses the importance of a population health approach and points out the need to improve access to high-quality, cost-effective health care in the United States. *Crossing the Quality Chasm* proposes six aims for improvement to address six dimensions of health care: safety, effectiveness, patient-centeredness, timeliness, efficiency, and equitableness (Institute of Medicine of the National Academies, 2001). The *2004 National Healthcare Quality Report* provides an update on the nation's progress in transforming health care in relation to these six aims and delineates major data sources available to track cost and quality outcomes in relation to population health. Many of the data sources being used to track *Healthy People 2010* objectives overlap with sources used to track *National Healthcare Quality* measures. But some sources such as the Healthcare Cost and Utilization Project (HCUP) data set and Surveillance, Epidemiology, and End Results Program data are being used only to track quality measures (AHRQ, 2004).

There are abundant clinical, administrative, and health care survey data sets available for population health. Current national health imperatives, increased emphases on population health, and ready availability and retrievability of existing data provide the perfect setting to increase nursing's involvement with health services research and analysis of population health data.

3. Issues associated with secondary analysis of population health data

Bierman and Bubolz (2003), Clarke and Cossette (2000), Hearst, Grady, Barron, and Kerlikowske (2001), Pollack (1999), and Smaldone and Connor (2003) point out similar limitations associated with the use of secondary data analysis. These limitations can be summed up into three main categories, in relation to analysis of population health data: (1) difficulty in locating required data, (2) incongruity of primary and secondary research objectives, and (3) data quality.

3.1. Locating existing data

Difficulty in locating existing data can be resolved in a number of ways. Data can be located through searching organization, survey, and data warehouse web sites and through the review of published research studies (Clarke & Cossette, 2000; Smaldone & Connor, 2003).

3.1.1. Data available through organization and survey web sites

Table 1 depicts some of the major publicly available administrative, clinical, and national survey data sets. The National Center for Health Statistics (NCHS) web site (http://www.cdc.gov/nchs/Default.htm) provides links to several Centers for Disease Control and Prevention (CDC) national survey and data collection system data sets. The HHS web site (http://www.os.dhhs.gov/) provides links to several HHS agencies that provide health services research funding opportunities, as well as public-use access to administrative, clinical, and national health survey data sets. Privacy-protected data can be obtained after a data use agreement has been completed and processed, but public-use, no-cost data can be accessed directly from the organization, agency, or survey web site.

3.1.2. Data available through data warehouse web sites

The Inter-University Consortium for Political and Social Research (ICPSR) web site (http://www.icpsr.umich.edu/) links to one example of a data warehouse for existing administrative, clinical, and national health survey data (Inter-University Consortium for Political and Social Research (ICPSR), 2005). The ICPSR maintains and provides access to a huge archive of social science data for research

Table 1
Selected major sources of publicly available clinical, administrative, and national health survey data from the United States

Major data source, sponsors, and home page web sites	Overview of data source
Behavioral Risk Factor Surveillance System (BRFSS), http://www.cdc.gov/brfss/ Sponsor: CDC	Data collected in 50 states, three territories, and three U.S. territories since 1994. Annual data release. BRFSS is a cross-sectional telephone survey conducted each year by state health departments with technical and methodological assistance provided by the CDC. States conduct monthly telephone surveillance using a standardized questionnaire to determine the distribution of risk behaviors and health practices among noninstitutionalized adults within the states. States forward the responses to the CDC, where monthly data are aggregated for each state. The data are returned to the states and published on the BRFSS web site. Public-use data are available, without cost. See web site.
HCUP, http://www.ahrq.gov/data/hcup/ Sponsors: HHS and AHRQ	Data collected since 1988. Annual data release. HCUP is a collection of longitudinal hospital care data, with all-payer, discharge-level information. The Nationwide Inpatient Sample includes inpatient data from a national sample of more than 1,000 hospitals; the State Inpatient Databases cover inpatient care in community hospitals in participating states that represent about 90% of all U.S. hospital discharges. Population targeted: U.S. citizen or foreign, using nonfederal, community hospitals in the United States. Access to an online query system based on HCUP databases is available free of charge. Many of the databases featured in HCUP can be purchased for the purpose of performing individual analyses. See web site.
Medical Expenditure Panel Survey (MEPS), http://www.meps.ahrq.gov/ Sponsors: HHS, AHRQ, CDC, and NCHS.	Data collected since 1996. Annual data release. MEPS is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. Data include demographic characteristics, health conditions, health status, use of medical care services, payments, access to care, satisfaction with care, health insurance coverage, income, and employment. Population targeted: nationally representative survey of the U.S. civilian noninstitutionalized population. Public-use data are available, without cost. See web site.
Minimum Data Set (MDS), http://www.cms.hhs.gov/MinimumDataSets20/ Sponsors: HHS and Centers for Medicare & Medicaid Services (CMS)	Data collected since 1990 by nursing homes. Since June 1998, states have transmitted data to CMS. The MDS is a standardized, primary screening and assessment tool and measures physical, medical, psychological, and social functioning of nursing home residents. Data are collected by facility personnel, with attestation of accuracy and transmitted electronically to states. Population targeted: all residents in Medicare- or Medicaid-certified nursing and long-term care facilities. Long-Term Care MDSs are classified as identifiable and are available with an associated cost. Requests for Long-Term Care Minimum Care Data Set data must be submitted to the Research Data Assistance Center (ResDAC), at http://www.resdac.umn.edu. Once data requests are reviewed, these requests can be forwarded to CMS for processing, as directed by ResDAC.
National Health Interview Survey, http://www.cdc.gov/nchs/nhis.htm Sponsors: HHS, CDC, and NCHS	Conducted continuously since 1957. Annual data releases. Personal interview in households. Uses stratified multistage probability design. Population targeted: civilian noninstitutionalized population of all ages residing in the United States. Information obtained includes demographic characteristics, illnesses, injuries, impairments, chronic conditions, utilization of health resources, health insurance, and other health topics. Public-use data are available, without cost. See web site.
National Health and Nutrition Examination Survey, http://www.cdc.gov/nchs/nhanes.htm Sponsors: CDC and NCHS	From 1960 to 1994, a total of seven national examination surveys were conducted. From 1999 onward, surveys were conducted continuously with data releases every 2 years. Uses a stratified multistage probability sample and in-person interviews in the household and in a private setting in the mobile examination center. Population targeted: the civilian noninstitutionalized population residing in the United States aged 2 months and above. Beginning in 1999, people of all ages were included. Information obtained includes chronic disease prevalence and conditions, risk factors, diet and nutritional status, immunization status, infectious disease prevalence, health insurance, and measures of environmental exposures. Public-use data are available, without cost. See web site.
National Immunization Survey, http://www.cdc.gov/nis/ Sponsors: CDC, NCHS, and the National Immunization Program	Conducted quarterly since 1994. A list-assisted random-digit-dialing telephone survey, conducted in each of 78 Immunization Action Plan areas (which together make up the United States). Population targeted: children between the ages of 19 and 35 months living in the United States at the time of the interview. Information obtained includes vaccination coverage rates for each of six recommended vaccines for the United States. Public-use data are available, without cost. See web site.
National Vital Statistics System—Mortality, http://www.cdc.gov/nchs/deaths.htm Sponsors: CDC and NCHS	Data system began in 1880 but all states did not participate before 1933. Since 1933, coverage for deaths has been complete. Annual data releases include information from mortality files for the 50 states, the District of Columbia, and the territories of Puerto Rico, Virgin Islands, Guam, American Samoa, and the Commonwealth of the Northern Marianas. Data are obtained through administrative records (death certificates) completed by funeral directors, physicians, medical examiners, and coroners and filed with state vital statistics offices. Information obtained includes year of death, place of decedent's residence, place where death occurred, age at death, day of week and month of death, Hispanic origin, race, marital status (beginning in 1979), place of birth, gender, underlying and multiple causes of death for all states, injury at work (beginning in 1993), hospital and patient status, and educational attainment (beginning in 1989) for selected states. Some aggregate data files are available, without cost. Other data files are available with an associated cost. See web site.
Surveillance, Epidemiology, and End Results Program, http://seer.cancer.gov/resources/ Sponsor: National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program, Cancer Statistics Branch	Data collected since 1973. Annual release of data collected and published on cancer incidence and survival from 12 population-based cancer registries and 3 supplemental registries covering approximately 14% of the U.S. population. Population targeted: age-adjusted U.S. population (see web site for more information). Data include patient demographics, primary tumor site, morphology, stage at diagnosis, first course of treatment, and follow-up for vital status. Public-use data are available, without cost. See web site.

Note. Table content was adapted from specific web sites for the major data sources included in the table.

and instruction. ICPSR data sets are housed under several thematic categories, and population-health-oriented data can be accessed by browsing the "Health Care and Facilities" category. Most data can be retrieved free of charge, after applying for a password and completing an online data use agreement. Some data sets are available only to individuals from colleges and universities maintaining membership with ICPSR. A list of organizations can be viewed on the ICPSR web site. Although many of the national health survey data sets available through ICPSR can be obtained directly from the national health survey organization's web site, ICPSR provides access to some data sets that are not otherwise publicly available.

3.1.3. Data available through the review of published studies

Sources of clinical, administrative, and survey data can be located by conducting a search of the literature to identify published population-health-related research studies. Studies can be located by using specific key words relating to a particular disease condition or health indicator such as "diabetes" or "physical activity" in combination with the term "secondary analysis," or a broader search can be conducted using the key words "population health" and "secondary analysis." The types of results returned vary across bibliographic databases, even when the same key words are used. For instance, in a search of the literature conducted using the key words "population health" and "secondary analysis" to identify sources of population health data in published studies relating to populations within the United States between January 1, 2004, and August 30, 2005, approximately 60 studies were located through the PubMed bibliographic database as compared to 0 studies located through the Cumulative Index of Nursing and Allied Health Literature (CINAHL). When a similar search was conducted using the same search criteria with the key words "secondary analysis" and "mental health," 11 studies were located through CINAHL and 8 studies were located through PubMed. There was very little overlap in the studies found in both bibliographic databases, and there was no overlap in the studies located that were published in nursing journals. Of the 11 studies located through CINAHL, 2 were published in nursing journals; of the 8 studies located through PubMed, 1 was published in a nursing journal.

The search results reported above are final numbers of studies retrieved that met the concept definitions of population health and secondary analysis presented earlier in this article. The studies retrieved also excluded all studies based on population health data for populations outside the United States. When searching for sources of existing data in published studies, one must clearly define key word concepts and population group parameters before beginning the search. In addition, searches should be conducted in the nursing, allied health, health services, and medical literature and employ the use of several bibliographic databases.

3.2. Development of research questions

Incongruity of primary and secondary research objectives is inherent to the secondary data analysis methodological design. According to Hearst et al., (2001), this specific limitation can be addressed in one of two ways: finding research questions to fit existing data or find existing data to fit specific research questions.

3.2.1. Finding research questions to fit existing data

The first step in finding a research question to fit existing population health data is to choose a specific database to draw the data from. After the database is selected, one should become familiar with the database and develop a flow sheet of the variables contained in the data set (Hearst et al., 2001). The next step in this process is to identify pairs of variables or relationships among the variables that might be of interest and conduct a review of the literature to establish the relevance and significance of exploring the relationships. Once the need for study of the relationships has been justified, the research question(s) and analysis plan can be developed. Data can then be accessed, and analyses can then be conducted.

3.2.2. Finding existing data to fit a specific research question

Location of existing data to fit a specific research question begins with the formulation of the question. Once the question is developed, one must conduct a thorough review of the literature to construct combinations of predictor and outcome variables that might help answer the question. The next step is to locate databases that might include the variables of interest (Hearst et al., 2001). These databases can be located using one of the methods discussed earlier in this article. The final steps in this process include choosing the best database for the research question, developing the analysis plan, accessing the data, and conducting the analyses.

In both instances discussed above, one should become familiar enough with the characteristics associated with the design, development, and quality of the data set to construct a proposed analysis plan. This understanding can emanate from reviewing codebooks, summary reports, and background information related to the data set; from reviewing previously conducted studies that used secondary analysis of the data set as a research methodology; and from talking with individuals familiar with the data set. In some cases, there may be an opportunity to review the actual data. However, issues related to data use agreements, research plan approval, and costs associated with retrieving the data may require development of the proposed analysis plan before data can be accessed.

3.3. Data quality

Although it is not possible to participate in deciding which data will be collected, when data already exist, one

can become familiar with characteristics associated with the existing data. Familiarity with the data set and access to information related to the purpose, content, development, population representation, and previously conducted studies relating to the existing data will provide one with the information required to assess the quality of the data. Pollack (1999) discusses quality of secondary data in terms of reliability and validity of the data. According to Pollack, reliability addresses how data were collected and coded for entry into the database. Specifically, reliability speaks to arrangement and accumulation of the data in an accurate and consistent manner that is replicable, whereas validity is concerned with the degree to which the data set contains all of the variables required to address the research questions being examined.

Statistical analyses of secondary data should be approached in the same manner as statistical analyses of primary data. Data should be organized and cleaned, and frequencies should be conducted to check the quality of the data entered into statistical programs before conducting analyses. However, there are issues specific to the quality of secondary data that should be addressed before one decides to use a particular data set to address a set of secondary research questions. Supporting documentation such as codebooks, summary reports, research proposals, and published studies should be available to conduct an initial assessment of the completeness and accuracy of the data set. Content of the documentation should be sufficient enough to address the following questions, at a minimum: Who is the sponsor/collector/owner of the data? What is the purpose of the study, survey, or development of the database? What is the research design? How were data collected? What sampling procedures were used? What is the number of observations? What are the variables? How many data are missing? How are the data coded? Are summary statistics available?

Existing data sets sponsored by and associated with the HHS are accepted rich sources of data on the health status and health needs of the U.S. population. The information pertaining to the completeness and accuracy of these data sets is readily available through survey and organization web sites, as well as through data warehouse archives. Also, data made available through data warehouses such as ICPSR come with supporting documentation and a list of related literature and published studies. However, data sets available from nonstate or nonfederal sources, from unpublished primary studies, and from local health care organization clinical, administrative, and health survey databases may not have sufficient supporting documentation to appropriately assess the quality of a data set. If supporting documentation is not available or does not contain sufficient information to conduct this initial assessment, one should consider using a different data set. If a decision is made to acquire and use the data set for research, without sufficient documentation to evaluate completeness and accuracy, limitations associated with invalid or unreliable data should be clearly addressed in the research report.

4. Conclusion

The current health care requirements necessitate that nursing practice be evidence based, population health oriented, and outcome driven. Secondary analysis of population health data is an excellent mechanism for nurses to use to increase the knowledge base required to support the demands of nursing practice in the 21st century, while increasing nursing's involvement with population health and health services research. Secondary analysis is a practical, cost-effective, and scientific method for gaining access to population health data but is associated with issues that can impact expediency, economy, and reliability. These issues can and must be addressed because use of poor-quality, inappropriate, unrepresentative data can lead to inaccurate conclusions and affect the reliability and validity of evidence generated to support current practice.

Acknowledgments

The views expressed in this article are those of the author and do not reflect the official policy or position of the Uniformed Services University of the Health Sciences, the Department of Defense, or the U.S. Government.

The author would like to thank Barbara Sylvia, PhD, RN, Professor and Acting Chair, Department of Health, Injury, and Disease Management, Graduate School of Nursing Uniformed Services University of the Health Sciences, for reading the original version of the manuscript.

References

Agency for Healthcare Research and Quality [AHRQ]. (2004). *National healthcare quality report*. Available at 2004 National Healthcare Quality Report Website http://www.qualitytools.ahrq.gov/qualityreport/browse/browse.aspx.

Bierman, A., & Bubolz, T. (2003, October). *Secondary analysis of large survey databases*. In M. B. Max, J. Lynn (Eds.), *Interactive textbook on clinical symptom research* (chap 20, pp. 2–3) Available at: http://symptomresearch.nih.gov/chapter_20/index.htm.

Bowling, A. (2002). *Research methods in health: Investigating health and health services,* (2nd ed). Maidenhead: Open University Press.

Castle, J. E. (2003). *Maximizing research opportunities: Secondary data analysis*. *Journal of Neuroscience Nursing, 35*, 287–290.

Clarke, S. P., & Cossette, S. (2000). Secondary analysis: Theoretical, methodological, and practical considerations. *Canadian Journal of Nursing Research, 32*, 109–129.

Fos, P. J., & Fine, D. J. (2005). *Managerial epidemiology for health care organizations*, (2nd ed). San Francisco: Jossey-Bass.

Hearst, N., Grady, D., Barron, H. V., & Kerlikowske, K. (2001). Research using existing data: Secondary data analysis, ancillary studies, and systematic reviews. In S. B. Hulley, S. R. Cummings, W. S. Browner, D. Grady, N. Hearst & T. B. Newman (Eds.), *Designing clinical research*, (2nd ed). Philadelphia, PA: Lippincott Williams and Wilkins.

Hughes, R. G. (2004). Some tips on getting funding for health services research. *Applied Nursing Research, 17*, 305–307.

Institute of Medicine of the National Academies [IOMNA]. (2001). *Crossing the quality chasm: A new health system for the 21st century.* Available from Institute of Medicine Website http://www.iom.edu/.

Inter-University Consortium for Political and Social Research (ICPSR). (2005). *Data access and analysis*. Accessed September 10, 2005 at http://www.icpsr.umich.edu/access/index.html.

Jones, C. B., & Lusk, S. L. (2002). Incorporating health services research into nursing doctoral programs. *Nursing Outlook, 50*, 225–231.

Kindig, D., & Stoddart, G. (2003). What is population health? *American Journal of Public Health, 93*, 380–383.

Pollack, C. D. (1999). Methodological considerations with secondary analyses. *Outcomes Management for Nursing Practice, 3*, 147–152.

Smaldone, A. M., & Connor, J. A. (2003). The use of large administrative data sets in nursing research. *Applied Nursing Research, 16*, 205–207.

U.S. Department of Health and Human Services [HHS]. (2000). *Healthy people 2010*. Available from Healthy People 2010 Website http://www.healthypeople.gov.

U.S. Department of Health and Human Services [HHS]. (2005). *Healthy people 2010: Data overview*. Retrieved September 10, 2005 from http://www.healthypeople.gov.

Vogt, W. P. (2005). *Dictionary of statistics and methodology,* (3rd ed.). Thousand Oaks: Sage.