Healthcare Survey Analytic and Data Quality Enrichments Achieved Through Consolidation

Steven B. Cohen
RTI International, P.O. Box 12194, Research Triangle Park, NC 27709-2194

Abstract

The quality and content of national population-based socioeconomic and health care surveys are enhanced through linkage to surveys of associated medical providers, businesses, and facilities. Analytical capacity is dramatically enhanced through their connectivity to existing secondary data sources at higher levels of aggregation and via direct matches to additional health and socioeconomic measures acquired for the same sample units from other sources of survey or administrative data. These administrative databases also may serve as sampling frames to facilitate a cost-efficient sample selection. These designs improve data collection strategies to meet target response rates, achieve reductions in nonresponse bias, and enhance data quality and analytical capacity. They permit extensions in longitudinal analyses and permit methodological studies to assess the accuracy of household reported data. Advances in data science also serve to facilitate the effective and efficient utilization of statistical models and procedures in concert with big data applications. The design features and analytic enhancements are illustrated with examples drawn from national health-related surveys with coalesced designs. They include the Medical Expenditure Panel Survey (MEPS), the MEPS Medical Provider Component (MPC) and Medical Organization Survey (MOS), and the Health Interview Survey (NHIS). Design limitations also are discussed.

Keywords: Consolidated survey designs, data quality, healthcare analytics

1. Introduction

The quality and content of national population-based socioeconomic and health care surveys are enhanced through linkage to surveys of associated medical providers, businesses, and facilities. Analytical capacity is dramatically enhanced through their connectivity to existing secondary data sources at higher levels of aggregation and via direct matches to additional health and socioeconomic measures acquired for the same sample units from other sources of survey or administrative data. These administrative databases also may serve as sampling frames to facilitate a cost-efficient sample selection. These designs improve data collection strategies to meet target response rates, achieve reductions in nonresponse bias, and enhance data quality and analytical capacity. They permit extensions in longitudinal analyses and permit methodological studies to assess the accuracy of household reported data. Advances in data science also serve to facilitate the effective and efficient utilization of statistical models and procedures in concert with big data applications. The design features and analytic enhancements are illustrated with examples drawn from national health-related surveys with coalesced designs. They include the Medical Expenditure Panel Survey (MEPS), the MEPS Medical Provider Component (MPC) and Medical Organization Survey (MOS), and the Health Interview Survey (NHIS). Design limitations also are discussed.

2. Analytical Enhancements Achieved Through Linkage of Surveys to Other Sources of Data

The analytical capacity of health surveys can be dramatically enhanced through the linkage to existing secondary data sources at higher levels of aggregation (both geographic and organizational) as well as through direct matches to additional health and socio-economic measures acquired for the same set of sample units from other sources of survey specific or administrative data. One of the more pervasive uses of existing administrative data bases is to serve as a sampling frame to facilitate a cost efficient identification of an eligible survey population for purposes of sample selection, such as the consideration of the Medicare administrative records to serve as a sampling frame for a survey of Medicare beneficiaries. Health surveys that are so linked to administrative records from their inception benefit by this capacity for data supplementation that permits enhanced and more extensive analyses that are beyond the more constrained scope of the core health survey. Establishing similar connections to existing data sources that will substantially enhance a survey's capacity to address specific research questions is often more difficult to establish after a survey has been administered. This is primarily a consequence of confidentiality restrictions that require respondent permission to link patient records to administrative data sources, in addition to problems with the availability of the necessary identifiers from the survey respondents [9].

The large majority of the nationally representative population-based health surveys sponsored by the Department of Health and Human Services have benefited by a capacity to link the survey data to county level data on health service resources and health manpower statistics available on the Area Health Resources File (AHRF). More specifically, the AHRF is a county-specific health resources information system containing information on health facilities, health professions, measures of resource scarcity, health status, economic activity, health training programs, and socio-economic and environmental characteristics. Geographic codes and descriptors are provided to enable linkage to health surveys to expand analyses conducted by planners, policymakers, researchers, and other professionals examining the nation's health care delivery system and in factors that may impact health status and health care in the U.S. Comparable enhancements to health surveys for supplementation of economic indicators are achievable through linkage of survey data to the socio-economic indicators made available by the Bureau of the Census through the County and City Data Book and public use files from the decennial Census.

The quality and data content of household specific health surveys are often enhanced through the conduct of follow back surveys to medical providers and facilities that have provided care to household respondents. In terms of data quality, household reported medical conditions can be evaluated for accuracy relative to provider specific records on medical conditions for the same patient and specific health events. With respect to health care expenditures collected from household respondents for their reported health care events, available linked medical provider level data is a more accurate source of information. The availability of such supplemental data on use and expenditures allows for the conduct of methodological studies to evaluate the accuracy of household reported data and informs adjustment strategies to household data in the absence of provider specific data to reduce bias attributable to response error [9].

3. Applications to the Medical Expenditure Panel Survey

One of the core health care surveys in the United States, the Medical Expenditure Panel Survey (MEPS), is characterized by a consolidated survey design. Since its inception, the primary analytical focus of the MEPS has been directed to the topics of health care access, coverage, cost and use. Over the past several years, the MEPS data have supported a highly visible set of descriptive and behavioral analyses of the U.S. health care system. These include studies of the population's access to, use of, and expenditures and sources of payment for health care; the availability and costs of private health insurance in the employment-related and non-group markets; the population enrolled in public health insurance coverage and those without health care coverage; and the role of health status in health care use, expenditures, and household decision making, and in health insurance and employment choices. As a consequence of its breadth, the data have informed the nation's economic models and their projections of health care expenditures and utilization. The level of the cost and coverage detail collected in the MEPS has enabled public and private sector economic models to develop national and regional estimates of the impact of changes in financing, coverage, and reimbursement policy, as well as estimates of who benefits and who bears the cost of a change in policy.

The Medical Expenditure Panel Survey (MEPS) has been collecting data on health care utilization and expenditures annually since 1996. The survey is sponsored by the Agency for Healthcare Research and Quality (AHRQ). In addition to collecting nationally representative data to yield annual estimates for a variety of measures related to health care use and expenditures, the MEPS also provides estimates related to health status, demographic characteristics, employment, health insurance coverage, and access to health care. The MEPS consists of a family of three interrelated surveys: the Household Component (MEPS-HC), the Medical Provider Component (MEPS-MPC), and the Insurance Component (MEPS-IC). The MEPS-IC also collects establishment-level data on insurance programs. Through a series of interviews with household respondents, the MEPS-HC collects detailed information at the level of the individual respondent on demographic characteristics, health status, health insurance, employment, and medical care use and expenditures. These data support estimates both for individuals and for families in the United States. Respondents identify medical providers from whom they have received services [1-10].

The set of households selected for the Household Component is a subsample of those participating in the National Health Interview Survey (NHIS), an ongoing annual household survey of approximately 40,000 households conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention, to obtain national estimates of health care utilization, health conditions, health status, insurance coverage and access. In addition to the cost savings achieved by eliminating the need to independently list and screen households, selecting a subsample of NHIS participants has resulted in an enhancement in analytical capacity of the resultant survey data. Use of the NHIS data in concert with the data collected for the MEPS provides an additional capacity for longitudinal analyses not otherwise available. Furthermore, the large number and dispersion of the primary sampling units in MEPS has resulted in improvements in precision over prior expenditure survey designs. The MEPS HC survey consists of an overlapping panel design in which any given sample panel is interviewed a total of 5 times in person over 30 months to yield annual use and expenditure data for two calendar years. These rounds of interviewing are spaced about 5 to 6 months apart. The interview is administered through a computer assisted personal interview mode of data collection, and

takes place with a family respondent who reports for him/herself and for other family members. Data from two panels are combined to produce estimates for each calendar year.

The MEPS Medical Provider Component is a survey of the medical providers, facilities and pharmacies that provided care or services to sample persons. The primary objective is to collect detailed data on the expenditures and sources of payment for the medical services provided to individuals sampled for the MEPS. Such data are essential to improve the accuracy of the national medical expenditure estimates derived from the MEPS, since household respondents are not always the most reliable source of information on medical expenditures.MPC data are collected a year after the household health care event information is collected to allow adequate time for billing transactions to be completed. The MPC collects data on dates of visits/services, use of medical care services, charges, sources of payments and amounts, and diagnoses and procedure codes for medical visits/encounters. Only providers for whom a signed permission form was obtained from the household authorizing contact are eligible for data collection in the MPC [9,10]. The categories of providers in the MPC include (1) office-based medical doctors; (2) hospital facilities providing inpatient, outpatient, and emergency room care; (3) health maintenance organizations (HMOs); (4) physicians providing care during a hospitalization; (5) home care agencies; and (6) pharmacies. RTI International is the data collection organization for the MEPS MPC.

In 2016, a linked Medical Organization Survey (MEPS-MOS) was added to the MEPS. The principal objectives of this MEPS design enhancement were (1) to develop procedures for identifying the medical organizations associated with the usual source of office-based ambulatory care physicians from whom a nationally representative sample of individuals receive medical care; (2) to refine a survey questionnaire designed for assessing important features of the staffing, organization, policies, and financing of office-based and related ambulatory care medical care providers; (3) to collect organizational level data associated with these providers of medical care to MEPS respondents; (4) to develop estimation weights that support nationally representative linked provider-respondent data based on the MEPS-MOS survey; and (5) to make the linked provider-respondent data set available to the research community.

Since the MEPS survey currently does not acquire essential data on providers, practice and organizational characteristics, policies and treatment protocols, penetration of ACOs, medical homes and health information technology (HIT), this survey will fill a critical gap in content. The following areas will be addressed in the MOS survey as they potentially affect individuals' access to, use of, and affordability of health care services:

- Organizational characteristics, e.g., size, specialties covered, practice rules and procedures, patient mix and scope of care provided, membership in an ACO, certification as a primary care medical home
- Use of health information technology
- Policies and practices related to the ACA
- Financial arrangements, e.g., reimbursement methods, number and types of insurance contracts, compensation arrangements within the practice

The enhanced MEPS data will benefit health care policymakers and health services researchers primarily by filling a gap in the available evidence linking provider characteristics with individual behavior and outcomes. Accurate and timely information on physicians' practices medical organizations is essential to understanding the functioning of the health care system, identifying potential problems, and assessing programmatic and policy reforms. Recent health reform initiatives are attempting to stimulate health care system improvements through advancing the role of primary care physicians in care management, providing incentives to physicians for the delivery of high-quality care and facilitating delivery system transformations to improve care, with special attention to the treatment of patients with multiple chronic conditions. Physicians and their practice organizations are integral to these initiatives, so understanding the organizational context in which they practice and how they practice and respond to policy and economic incentives is a critical input for evaluating and predicting the success of such reforms. Consequently, this design modification will help advance research efforts to discern how recent changes in health care delivery and practice resulting from the Affordable Care Act's health reform efforts affect health care costs, access, health status and health care quality.

The Medical Expenditure Panel Survey Insurance Component (MEPS-IC) is a nationally representative annual survey of over 40,000 business establishments and state/local governments. The survey is designed to produce estimates at the national and State level on the number and types of private health insurance plans offered, benefits associated with these plans, premiums, contributions by employers and employees, eligibility requirements, and employer characteristics. The cross-sectional MEPS-IC design provides estimates of employer decisions about health insurance offerings prior and post full implementation of the coverage provisions in the Affordable Care Act, both at the national and State level. While cross-sectional surveys permit analyses of net changes in population parameters at an aggregate level, only a longitudinal survey can discern the extent to which this is attributable to different elements of gross change. For example, under a cross-sectional design, consider a situation in which the annual estimates of employer health insurance offer rates were estimated to be the same over two consecutive years. Only a longitudinal design could determine whether it was the same set of employers who maintain their coverage offer rates or whether there were substantial counter-balancing shifts in employer sponsored coverage over time. Recently, the survey has also added a longitudinal arm to interpret direct changes in employer behavior over time.

4. Gains in Precision from a Longitudinal Design

In addition to the analytical attractions of a longitudinal design to assess changes in health behaviors over time, the use of each individual as its own control in analyses of time trends has additional benefits in terms of gains in precision using paired comparisons. To illustrate this expected gain in precision for analyzing changes in health care related behaviors through a design modification to allow for longitudinal analyses, the following analysis was conducted based on the MEPS Household Component, which has a longitudinal design. The following estimates derived from the survey for calendar years 2009 and 2010 were identified: annual healthcare expenditures, annual out of pocket healthcare expenditures, annual number of hospital stays, annual number of Dr. visits, the percent with fair/poor health status, and the percent of the population uninsured throughout the entire year. The sample was further restricted to those individuals who were classified as respondents for both years under study. The standard errors of the mean differences in survey estimates over the two years were then analyzed under two alternative survey design

assumptions: 1) the samples for each of the survey years were independently selected, and 2) the sample observations were obtained from a longitudinal survey design [10].

Table 1 provides a summary of the respective estimates of the standard errors of the mean differences in survey estimates for the specified health care measures under the two design options. The results clearly indicate that the standard errors obtained from a design with two independent sample selections for the two year period are consistently higher than those obtained from a longitudinal design, ranging from 1.17 to 2.24 times as large. However, it is important to note that a longitudinal design is often characterized by lower survey response rates for subsequent years post the initial contact relative to cross-sectional design as a consequence of survey attrition. Survey estimates under longitudinal designs are also subject to potential bias due to conditioning effects over time. Consequently, a decision regarding the optimal design for a given survey is often based upon weighing the competing benefits and limitations of the alternative designs under consideration.

Table 1: Comparison of Precision in Estimates Under Alternative Design Assumptions

Measure	Mean Difference over time (2010-2009)	Standard error – Independent Design	Standard error – Longitudinal Design	Ratio of S.E.s Independent Design/Longitudinal Design
annual healthcare expenditures	**69.4815**	**115.79543**	**91.23015**	**1.26927**
annual out of pocket healthcare expenditures	**56.9348**	**15.70015**	**13.30513**	**1.18001**
annual number of hospital stays	**-0.0021**	**0.00437**	**0.00375**	**1.16521**
annual number of Dr. visits	**0.1437**	**0.05914**	**0.03694**	**1.60095**
percent with fair/poor health status	**0.5431**	**0.36504**	**0.25401**	**1.43713**
percent of the population uninsured	**-0.5116**	**0.41347**	**0.18420**	**2.24472**

Source: Medical Expenditure Panel Survey 2009-2010, Agency for Healthcare Research and Quality

5. Summary

A key feature of a consolidated survey design is the direct linkage between sample members in the core survey with the larger host survey; administrative records; or follow-up surveys. In this paper, the capacity of integrated survey designs to achieve reductions in bias attributable to survey nonresponse is discussed. Several examples are drawn from the MEPS, which is linked to a host survey and has additional connections to follow-up surveys of medical providers and employers. In addition to utilizing this information as a frame to support the sample design of the core survey, this prior information from the host

survey or administrative records informs nonresponse and poststratification adjustments, imputation and serves as a data supplement for item nonresponse. The detailed information available on demographic/socio-economic characteristics of both respondents/and nonrespondents from the host survey or administrative records enhance the capacity of the specification of more direct nonresponse adjustments to better correct for survey nonresponse. In the absence of an integrated survey design, the nonresponse adjustment strategy adopted for surveys such as the MEPS would be constrained to socio-demographic and economic information that were available at the geographic level (e.g., county, state, division, and region).

A consolidated survey design model also provides additional features with respect to improving data collection strategies tied to the core survey to better ensure that target response rates are achieved. When the core survey is linked to a larger host survey, the survey operations and field staff that are armed with detailed record of calls data from the host survey will be better poised to commit and target necessary nonresponse conversion techniques to those cases that included reluctant or hard to reach respondents in the prior data collection effort. A consolidated survey design model offers enhancements to data quality and analytical capacity. It permits a cost efficient specification of a sampling frame for the core survey by utilizing an existing frame with detailed socio-demographic information to facilitate oversampling efforts and allow for dual frame designs. These features are in clear contrast to new frame construction and/or independent screening interviews that characterize unlinked survey design efforts. The design's capacity for data augmentation for a fixed time period, and the potential for longitudinal analyses over time through survey linkages are other attractive features of an integrated design framework. In health care surveys similar to the MEPS, the use of additional administrative data and medical records for survey participants permits additional methodological investigations and evaluations to examine the accuracy of household reported data. When differentials are observed in the response profiles through these evaluations and comparisons, the design permits well specified adjustment and estimation strategies to correct for measurement error [9].

It is important to note that several of the desired features of a consolidated survey design are the sources of its most prominent limitations. As a consequence of acquiring more information on survey respondents through data augmentation and data linkages over time, these analytical enhancements also increase the potential for disclosure of confidential information. To guard against this, it is necessary to impose greater restrictions on the release of data to the public. The sponsorship and operation of a data center to ensure that confidential data is in a secure environment while permitting more detailed analyses to be conducted with the non-publicly available data offers a compromise between greater data access and achieving confidentiality protection of data. However, this investment in the development and operation of a secure data center requires additional funds that may compete with sample size enhancements or planned research efforts.

A consolidated survey design also requires greater coordination across data sources and organizations. There are often competing demands on the host sample frames that may limit the full benefits of an integrated design from being realized. Furthermore, the enhanced longitudinal data that comes with an integrated survey design will often be characterized by more frequent survey contacts and rounds of data collection which will impact the overall survey response rate. When properly designed and coordinated, a consolidated survey design remains an attractive model for consideration and adoption [9].

Note: The views expressed in this paper are those of the author and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred. A substantial portion of the manuscript preparation was completed by Dr. Steven B. Cohen when he was Director of Center for Financing, Access and Cost Trends at the Agency for Healthcare Research and Quality.

References

[1]. R. Andersen 1968. A Behavioral Model of Families' Use of Health Services. Research Series #25. Chicago: Center for Health Administration Studies, University of Chicago.

[2]. R. Andersen and J. F. Newman, 1973. "Societal and Individual Determinants of Medical Care Utilization in the United States." *Milbank Memorial Fund Quarterly* 51 (Winter, 1973): 95-124.

[3]. J. Cohen, S. Cohen, and J. Banthin. The Medical Expenditure Panel Survey: A National Information Resource to Support Healthcare Cost Research and Inform Policy and Practice. *Medical Care* 47 (7, Suppl.1) (2009): 44–50.

[4]. S. B. Cohen and J. W. Cohen. The Capacity of the Medical Expenditure Panel Survey to Inform the Affordable Care Act. *Inquiry*. 50(2) (2013):124-134

[5]. S. B. Cohen and T. Buchmueller. Trends in Medical Care Costs, Coverage, Use and Access: Research Findings from the Medical Expenditure Panel Survey. *Medical Care* 44 (5) (2006): 1–3.

[6]. S.B. Cohen. Design Strategies and Innovations in the Medical Expenditure Panel Survey. *Medical Care* 41 (7) (2003): 5–12.

[7] Cohen S. B. Integrated Survey Designs: A Framework for Nonresponse Bias Reduction through the Linkage of Surveys, Administrative and Secondary Data. Agency forHealthcare Research and Quality Working Paper No. 04001, October 2004, http://www.ahrq.gov.

[8] Cohen SB, Cohen JW, Davis K. Longitudinal Design Options for the Medical Expenditure Panel Survey Insurance Component. Agency for Healthcare Research and Quality Working Paper No. 13003, May 2013, http://gold.ahrq.gov.

[9]. S. Machlin, S. and A. Taylor. Design, Methods, and Field Results of the Medical Expenditure Panel Survey Medical Provider Component. MEPS Methodology Report No.9. AHRQ Pub. No.00-0028. (2000). Rockville, MD.:Agency for Healthcare Research and Quality.

[10]. M. Stagnitti, K. Beauregard, and A. Solis. Design, Methods, and Field Results of the Medical Expenditure Panel Survey Medical Provider Component (MEPS MPC)—2006 Calendar Year Data, Methodology Report No. 23. November 2008. Agency for Healthcare Research and Quality, Rockville, MD.