Integrated Survey Designs: A Framework for Nonresponse Bias Reduction through the Linkage of Surveys, Administrative and Secondary Data

Steven B. Cohen

Agency for Healthcare Research and Quality Working Paper No. 04001

October 2004

Suggested citation: Cohen SB. Integrated Survey Designs: A Framework for Nonresponse Bias Reduction through the Linkage of Surveys, Administrative and Secondary Data. Agency for Healthcare Research and Quality Working Paper No. 04001, October 2004, http://gold.ahrq.gov.

AHRQ Working Papers provide preliminary analysis of substantive, technical, and methodological issues. The papers have not undergone external peer review. They are distributed to share valuable experience and research. Comments are welcome and should be directed to the authors. The views expressed are those of the authors and no official endorsement by the Agency for Healthcare Research and Quality or the Department of Health and Human Services is intended or should be inferred.

Integrated Survey Designs: A Framework for Nonresponse Bias Reduction through the Linkage of Surveys, Administrative and Secondary Data
Steven B. Cohen
October 2004

ABSTRACT

The quality and data content of household specific health surveys are often enhanced through integrated designs which include 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. The analytical capacity of surveys can also 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. In this paper, the capacity of integrated survey designs to achieve reductions in bias attributable to survey nonresponse is discussed. Examples are drawn from the Medical Expenditure Panel Survey (MEPS), an ongoing longitudinal panel survey designed to produce estimates of health care utilization, expenditures, sources of payment, and insurance coverage of the U.S. civilian non-institutionalized population.

Steven B. Cohen
Director, Center for Financing, Access, and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
E-mail: scohen@ahrq.gov

Integrated Survey Designs: A Framework for Nonresponse Bias Reduction through the Linkage of Surveys, Administrative and Secondary Data

Introduction

The quality and data content of household specific health surveys are often enhanced through integrated designs which include 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. The analytical capacity of surveys can also 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.

In this paper, the capacity of integrated survey designs to achieve reductions in bias attributable to survey nonresponse is discussed. Examples are drawn from the Medical Expenditure Panel Survey (MEPS), an ongoing longitudinal panel survey

designed to produce estimates of health care utilization, expenditures, sources of payment, and insurance coverage of the U.S. civilian non-institutionalized population.

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.

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 Resources File (ARF). More specifically, the ARF 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.

Applications to the Medical Expenditure Panel Survey

One of the core health care surveys in the United States, the MEPS, is characterized by an integrated 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.[1] 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.[2-7] 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.[8-10]

The MEPS consists of a family of three interrelated surveys: the Household

Component (HC), the Medical Provider Component (MPC), and the Insurance Component (IC). The survey is sponsored by the Agency for Healthcare Research and Quality (AHRQ). The MEPS Household Component was designed to provide annual national estimates of the health care use, medical expenditures, sources of payment and insurance coverage for the U.S. civilian non-institutionalized population. In addition to collecting data to yield annual estimates for a variety of measures related to health care use and expenditures, MEPS also provides estimates of measures related to health status, demographic characteristics, employment and access to health care. Estimates can be provided for individuals, families and population subgroups of interest. The data collected in this ongoing longitudinal study also permit studies of the determinants of the use of services and expenditures, and changes in the provision of health care in relation to social and demographic factors such as employment or income; the health status and satisfaction with health care of individuals and families; and the health needs of specific population groups such as the elderly and children.

Household Component

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 42,000 households (109,000 individuals) 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 (195 PSUs) in MEPS has resulted in improvements in precision over prior expenditure survey designs.[11]

The 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.[12] The initial year of the survey was 1996, and the household sample consisted of 8,655 families and 21,571 individuals with calendar year data . Currently, the MEPS sample consists of 15,000 families and 39,000 individuals, and reflects an oversample of the following policy relevant population subgroups: Hispanics, blacks, Asians and low income households. Data from two panels are combined to produce estimates for each calendar year.

Medical Provider Component

The 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. The data also serve as a primary imputation source of medical expenditure data to correct for the item nonresponse on this measure by the MEPS household sample participants.[13]

Medical providers (MD/DO) for whom household reported expenditure data was expected to be insufficient were sampled at higher rates. Households with one or more Medicaid enrollees and households with one or more persons enrolled in an HMO or managed care plan are oversampled because they were expected to have limited information about payments for their medical care. In addition, all hospitals providing inpatient and/or outpatient services to household members are contacted. The data collected from medical providers include: dates of medical encounters; medical content of each encounter, the charges associated with each encounter and the sources paying for the medical care. The data collected from pharmacies include: dates of prescription filled; prescription names; NDC codes; charges and payments by source. The data collection mode for hospitals, physicians and home health agencies was flexible, consisting of an initial telephone contact and then a mail or phone survey to collect specific information. The data collection mode for pharmacies was designed as a mail survey with telephone follow-up. In 2004, the Medical Provider Survey consists of interviews with more than 4,000 facilities, 22,000 office-based providers, 11,000 hospital-identified physicians, 800 home health providers and 9,000 pharmacies.

Insurance Component

The MEPS Insurance Component was designed to produce national and state

level estimates of the cost of employer sponsored coverage. National, regional, and State estimates can be made of the amount, types, and costs of job-related health insurance. Interviews are conducted annually via mail with 30,000 establishments to obtain national and state-specific estimates of the availability of health insurance at the workplace, the type of coverage provided by employers, and the associated costs of coverage. For each establishment surveyed (78% response rate), information was obtained on the number and characteristics of plans offered, the scope and breadth of benefits included in each plan and the corresponding co-payment provisions, the number of current workers and retirees enrolled in each plan, and whether each plan is fully or self-insured. The data collected also included characteristics of each establishment including its size, the type of workforce employed, aggregate data on payroll and available fringe benefits, industrial classification, and corporate status[14]. Comparable data were also collected for employers linked to the MEPS Household Survey. National estimates of employer health insurance premium costs obtained from this survey are now used by the Bureau of Economic Analysis to produce national estimates of the Gross Domestic Product as a consequence of the survey's data quality. The data are also being used to inform the national health care cost estimates in the National Health Accounts and to assess time trends in the provision of employer health benefits by states.

Integrated sample design in MEPS yields more efficient sample design

The original MEPS sample design called for an independent screening interview to identify a nationally representative sample and facilitate oversampling of policy-relevant population subgroups. Detailed information was to be obtained on socio-demographic,

economic and health status measures to support an oversample of the following policy relevant groups:

- Adults (18 years and older) with functional impairments.
- Children with limitations of activity.
- Individuals 18-64 years who were predicted to incur high medical expenditures.
- Individuals predicted to have family income less than 200 percent of the poverty level.

Detailed probabilistic models were to be used to target the oversample of individuals likely to incur high levels of expenditures in addition to those with family incomes less than 200 percent of the poverty level[15]. Data collection and training costs associated with this independent screening interview were projected to exceed $8 million. As part of the DHHS Survey Integration Plan, this separate screening interview was eliminated. Instead, NHIS was specified as the sampling frame for MEPS. In addition to the cost savings achieved by substituting NHIS as the MEPS sample frame, the design modification will result in an enhanced analytic capacity of the resultant survey data. Use of the NHIS data in concert with the MEPS data provides an additional capacity for longitudinal analyses not available in the original design. Furthermore, the greater number and dispersion of the sample primary sampling units that comprise the MEPS national sample should result in improvements in precision over the original design specifications.

Capacity to reduce bias attributable to survey nonresponse

As a consequence of the complex design of the MEPS HC, the MEPS sample data must be appropriately weighted to obtain approximately unbiased national estimates for the U.S. civilian noninstitutionalized population. The sampling weights developed for this purpose reflect the disproportionate sampling adopted in NHIS to oversample minority populations. They also reflect adjustments for:

- Complete nonresponse of eligible sample units.
- Partial response of survey participants providing data for only a portion of the time in 1996 during which they were eligible to respond.
- Poststratification to more accurate population totals obtained from the Current Population Survey (CPS).

The MEPS estimation weights are built from the estimation weights developed for the NHIS. Use of a sampling weight that has already incorporated the selection probabilities of the sample design and appropriate nonresponse and poststratification adjustments is an added feature of the integrated survey design. Since survey nonresponse is potentially a significant source of bias in survey estimates, the MEPS dwelling unit sampling weights included an adjustment to help reduce its potential for bias. In general, the greater the difference among subgroups in response rates and the analytic characteristic(s) of interest, the greater is the need to adjust survey weights for nonresponse. In MEPS, a weighting class nonresponse adjustment was implemented, under the assumption that nonresponding sampling units would have responded in a manner similar to that of respondents with similar sociodemographic and economic characteristics within the same

adjustment class. Properly designed, a weighting class nonresponse adjustment strategy can result in reduced nonresponse bias. The technique requires that the sample be partitioned into mutually exclusive classes, with classification information available for both responding and nonresponding units.[16] In the absence of an integrated survey design, the nonresponse adjustment strategy adopted for 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), rather than the detailed information available for each household participant in the NHIS sample selected for the MEPS. This is typical of standard household surveys which use aggregate data at the geographic level to inform the nonresponse adjustments (e.g., per capita income for the county based on secondary data available from the Census; physicians per 1,000 population and other health manpower statistics at the county level available from the Area Resources File).

Analyses were conducted of characteristics associated with differential nonresponse in the MEPS. These analyses identified the most important measures to use in developing a nonresponse adjustment to the MEPS sampling weights to correct for potential nonresponse bias at the dwelling-unit level. To facilitate these comparisons, the demographic, socioeconomic, health-related, and interview-specific profiles of respondents and nonrespondents were examined, based on available data for both groups from the NHIS.

Based on the results of these analyses, weighting classes were specified for the MEPS Round 1 dwelling unit nonresponse adjustments. They were defined by cross-classifications of the following measures:

- Family income of primary reporting unit (less than $10,000; $10,000-$19,999; $20,000-$34,999; $35,000 or more; unknown).
- Size of dwelling unit (one; two; three; four; five or more).
- MSA size (MSA, population 500,000 or more; MSA, population less than 500,000; non- MSA).
- Region (Northeast; Midwest; South; West).
- Employment classification of reference person (government; private sector; not in labor force/never worked/worked without pay; unknown or under 18 years of age).
- DU-level personal help measure (units with at least one member unable to perform personal care activities or other routine needs; remaining units with person 70 and over; remaining units with no limitations).
- Propensity to cooperate, based on providing phone number during NHIS (phone number provided; phone present but no number provided; no phone; unknown).
- Age of reference person (under 25; 25-34; 35-44; 45-64; 65 and over).
- Race/ethnicity of reference person (Hispanic; black non-Hispanic; other).
- Sex of reference person.
- Marital status (married, spouse present; other).

Overall, 49 cells were identified based on cross-classifications of these measures, with cell collapsing specified according to a hierarchy determined by significance level.

More specifically, the nonresponse adjustment for the cth weighting class takes the form

$$B(c) = \frac{\sum_{i \varepsilon c} E(i) DUPSWT(i)}{\sum_{i \varepsilon c} R(i) DUPSWT(i)}$$

where

DUPSWT(i) is the initial MEPS Round 1 dwelling unit weight for the ith sample dwelling unit, which reflects the reciprocal of the dwelling unit's selection probability for MEPS and a poststratification adjustment to 1995 NHIS population totals;

E(i) = 1 for all MEPS dwelling units selected for the Round 1 interview; E(i) = 0 otherwise;

R(i) = 1 for all selected MEPS dwelling units responding in Round 1, R(i) = 0 otherwise; and

iec represents eligible dwelling units classified in weighting class c.

Consequently, the estimation weight adjusted for MEPS Round 1 dwelling unit nonresponse, WGTDU1(i), for the ith dwelling unit associated with class c, takes the form

$$WGTDU1(i) = B(c) \times DUPSWT(i)$$

In the absence of an integrated survey design for the MEPS, none of the household specific information that were factors in the nonresponse adjustments would be available, other than the measures of MSA size and region. Clearly the MEPS linkage to the NHIS enhances the capacity of the specification of more direct nonresponse

adjustments to better correct for survey nonresponse.

Another survey that benefits by this integrated design model is the Medicare Current Beneficiary Survey (MCBS) sponsored by the Centers for Medicare and Medicaid Services. The MCBS is a continuous, multipurpose survey of a nationally representative sample of aged, disabled, and institutionalized Medicare beneficiaries. It provides a comprehensive source of information on the health status, health care use and expenditures, health insurance coverage, and socioeconomic and demographic characteristics of the entire spectrum of Medicare beneficiaries[18]. Rather than being linked to a larger survey, the sample for MCBS is drawn from administrative records in CMS's Medicare enrollment file. The Medicare enrollment files also provide mailing addresses for the sample. Medicare administrative files provide not only the sample frame but also service, diagnosis, and charge details for covered events, month-by-month information on enrollment status, payments for Medicaid buy-ins and HMO membership, and data for nonrespondents to the interview.

Linked Provider Data on Expenditures Improves the Accuracy of National Medical Expenditure Estimates in the MEPS

The MEPS Medical Provider Component (MPC) was primarily designed to reduce the bias associated with national medical expenditure estimates derived from household reported data. The estimation strategy that has been considered to support the data replacement strategy is comprehensive in nature, making full use of MPC data to correct for missing and poor quality household reported expenditure data. In addition, it

provides the basis for a recalibration of household reported data, if significant reporting differentials are observed in expenditure data between households and medical providers.

The foundation on which this estimation strategy rests is the household reported utilization experience. It is clearly recognized that household reports of medical utilization will be affected by errors of omission and commission that are a consequence of length of recall, memory loss, salience and proxy response. However, the primary focus of this estimation task is correct household expenditure estimates associated with a *household reported* medical event. At this stage in the MEPS estimation strategy, no adjustments to household reported utilization patterns are made. However, separate analyses are possible, using data on linked person-provider pairs, to assess the level of divergence between household and provider reports of health care utilization.

For the purposes of this estimation strategy, which combines the household reported and provider reported expenditure data, the unit of interest is the household reported utilization. A utilization event may be a visit to a specific doctor or clinic, or it may be an event involving several providers, such as a hospitalization. For a given calendar year, once the data collection phase of the MPC survey is completed, the first stage of this estimation strategy attempts to match all the provider reported expenditure data to the household reported utilization.[18-19]

For a sample person participating in the MPC, there are three distinct outcomes with respect to matching the MPC and the Household survey data. First, the household

respondent may report a utilization that matches to the data reported in the MPC. The second possibility is that utilization is reported in the MPC, but not by the person in the household survey. The third possibility is that a person may report a utilization that does not match any utilization in the MPC. This could happen if the permission form is not signed by the household respondent, if the provider does not respond to the MPC, if there is insufficient information to match their reports, if the provider did not give a complete response, or if the household respondent erroneously reported the event.

A computerized matching algorithm is then used to match household and provider reports of medical care utilization. The matching criteria include characteristics of the date of the utilization, the type of event (hospitalization, clinic visit, medical provider visit), and the household reported conditions and provider reported diagnoses that described the purpose of the utilization. The matching rules are developed to maximize the correct matches while minimizing the false matches and non-matches.

For all household and provider reported utilizations that match and for which MPC reported expenditure data exists (referred to as set A), the MPC data is used as the appropriate value of the expenditure:

Y_{ij} = MPC expenditure data for matched utilization j
associated with person i

For the subset of household and provider reported utilizations that match and for

which *both household and provider reported expenditure data* exist (referred to as subset A(l)), the relationship between these alternative sources of expenditure data is modeled to determine whether it is necessary to implement a recalibration procedure for cases with only household information. More specifically, let Y_{ij} be estimated as a model based function of X_{ij}, or

$$Y_{ij} = f(X_{ij}),$$

where

X_{ij} = HHS reported expenditure data for matched utilization *j*

associated with person *i*.

It is important to note that both Y_{ij} and X_{ij} are vectors of source of payment components which sum to the total expenditure and consist of self/family; Medicare; Medicaid; private; VA; Tricare; other federal; other state/local; workers compensation; other private; other public; remainder.

The purpose of a recalibration procedure is to rescale the person-reported data so that it is comparable to the provider reported data. The improvement from recalibration is based on the assumption that the provider's responses are more accurate than the person's expenditure responses. If it is determined that there are significant differentials

in the reporting patterns of medical expenditures between household respondents and their associated medical providers, the recalibration strategy should serve to reduce some of the bias in MEPS national expenditure estimates associated with person-level reporting.

Under this model, all remaining household reported utilizations not included in *A* for which a household reported expenditure is present, X_{ij} , (referred to as set B) would be recalibrated to a predicted provider reported response, using the model based function

$$Y_{ij} = f(X_{ij}).$$

Currently, none of the studies of comparisons of expenditure reports from the two distinct sources indicate that a recalibration adjustment is necessary. Since recalibration is not supportable, all remaining households not reported in *set A* for which household reported expenditure is present (set B) have their expenditure estimates specified as

$$Y_{ij} = (X_{ij}).$$

The remaining household reported utilizations not characterized in sets *A and B* for which no household reported expenditure data is present is corrected by a hot deck imputation strategy, utilizing the combination of replacement MPC and unadjusted household expenditure data that characterize the household respondents identified in sets *A* and *B.*

Integrated Design Features of the MEPS Facilitate Examination of Response Error: Options for Implementing an Adjustment to Household Reported Utilization Estimates Based on Provider Data

In addition to serving as the primary source for the expenditures in the MEPS, the design of the Medical Provider Component provides data that could potentially facilitate adjustments to household reported utilization data that correct for reporting errors (both under-reporting and over-reporting (telescoping errors)), under the assumption that the medical provider reports are the "gold standard". Within a given event type, the number of reported events were aggregated up to the person-provider pair level. The distribution of the difference in utilization counts between the medical provider and household reports was then examined. For each event type at the person-provider level (ij) , a difference measure, $DIFF_{ij}$, was computed, where:

$$DIFF_{ij} = MPSCOUNT_{ij} - HHSCOUNT_{ij}$$

$MPSCOUNT_{ij}$ = the number of events for the person-provider pair reported in provider survey, and

$HHSCOUNT_{ij}$ = the number of events for the person-provider pair reported in household survey.

Use of MPC data to develop adjustment factors that re-calibrate or correct

household reported data to reflect utilization counts based on MPC data offers a capacity to inform a utilization adjustment to correct for potential response error associated with household reports[20]. While the development of adjustment factors that correct for both under-reporting and over-reporting of health care utilization by household respondents is permissible, which would allow for household event counts to be either scaled down or up, based on reported or imputed MPS information, an alternative approach would be to limit the adjustment to correct the outlier cases (the poorest household reporters of utilization).

One evaluation of household reported and medical provider reported utilization data (matched at the person-provider pair level) revealed high levels of agreement between the two sources, and when there were differences, they were dominated by household under-reporting of use. For a given event type, the outlier cases could be identified by examining the ordered distribution of the utilization reporting agreement measure, DIFF_{ij}, and considering the point (the ordered value of DIFF_{ij}) at which the cumulative sum of the utilization reporting agreement measure crosses the value 0.0 as threshold point. For all linked person-provider pairs below this threshold, the household reporting errors of under-reporting and over-reporting of health care utilization are balanced in the aggregate.

Conditioned by event type, all linked person-provider pairs below the threshold would have their utilization adjustment factor, U_{ij}, specified as $U_{ij}=1$. These cases would be identified as $\mathrm{POOR_USE}_{ij}=0$.

All remaining linked person-provider pairs at or above the threshold would have their utilization adjustment factor, Uij, specified as

$$Uij = \text{MPSCOUNT}_{ij}/\text{HHSCOUNT}_{ij}.$$

These cases would be identified as POOR_USEij =1. If we wish to restrict the value of the upper bound of Uij, we could create a scale factor adjustment for the largest values of Uij (referred to as members of *c*), that will reduce the impact outlier scaling factors.

Here,

$$Uc = \sum_{ij \in c} \text{MPSCOUNT}_{ij} / \sum_{ij \in c} \text{HHSCOUNT}_{ij}.$$

Conditioned by event type, for the above subsets of linked person provider pairs, a logistic regression would be used to determine the factors associated with poor reporting, where POOR_USEij would be specified as the dependent variable. An analysis revealed that the number of household reported events, the length of recall, and education level were associated with the level of quality of household reported utilization. These measures would then be used to define the classification cells in order to implement a hot-deck imputation to allow for the specification of a utilization adjustment for the remaining person-provider pairs in the household data base. Furthermore, these linked person provider pairs would then serve as donor records for the hot-deck, where their assigned use adjustment factor, Uij, would be the measure to be imputed.

- Conditioned by event type, all remaining person-provider pairs would serve as recipients in a hot-deck imputation, and acquire the utilization adjustment factor, U_{ij}, based on their classification as a poor or good household reporter of utilization. The vast majority of recipients would acquire adjustment factors of U_{ij}=1. To insure that the resultant adjusted utilization measure was an integer value, the number of household reported visits would need to match exactly between donor and recipients (however, we could allow for non-integer values, particularly if we used the summary scale factor approach).

- At the person level, conditioned by event type, the adjusted utilization total, $USETOT_i$, would be obtained in the following manner:

$$USETOT_i = \sum_j U_{ij} x USE_{ij}$$

where USE_{ij} is the total number of household reported events for the given person-provider-pair ij, and U_{ij} is the utilization adjustment factor.

Summary

A key feature of an integrated 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 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).

In terms of the adoption of strategies to reduce the bias attributable to item nonresponse, the utility of an integrated survey design with a data replacement feature is an attractive approach, but necessitates the commitment of additional survey resources for required data collection and/or analytical matching and estimation tasks. When combined with “hot-deck”, “cold-deck” or model based imputation procedures to correct for remaining item nonresponse, an integrated survey design framework provides a more effective capacity to improve the accuracy of resultant survey estimates.

The integrated 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.

In addition to the gains to be achieved in the reduction of several major sources of nonresponse bias, an integrated 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.

It is important to note that several of the desired features of an integrated 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.

An integrated 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, as implemented for the MEPS, the integrated survey design remains an attractive model for consideration and adoption.

References

1. **Cohen JW, Monheit AC, Beauregard KM, et al.** The Medical Expenditure Panel Survey: A National Health Information Resource. Inquiry 1997;33:373-389

2. **Rhoades J, Vistnes J, Cohen, J**. The Uninsured in America: 1996-2000 AHRQ Pub. 2000:27

3. **Machlin S, Cohen J, Zuvekas, S et al.** Health Care Expenses in the Community Population. AHRQ Pub. 2001:27

4. **Holahan, J.** Health status and the cost of expanding insurance coverage. Health Affairs 2001; 20:279-286

5. **Blumberg L. Nichols L, Banthin J.** Worker decisions to purchase health insurance. Int. J Hth Fin Econ 2001;1:305-325

6. **Branscome J, Crimmel B**. Changes in Job-Related Health Insurance, 1996-99. AHRQ Pub. 2002:30

7. **Phillips K, Mayer M, Aday L.** Barriers to care among racial/ethnic groups under managed care. Health Affairs 2000;19:65-75

8. **Selden T, Banthin J, Cohen J.** Waiting in the Wings: Eligibility and Enrollment in the State Children's Health Insurance Program. Health Affairs 1999;18:126-133

9. **Cooper PF, Schone B**. More Offers, Fewer Takers for Employment-Based Health Insurance: 1987 and 1996. Health Affairs 1997:142-149.

10. **Selden TM, Moeller JF.** Estimates of the Tax Subsidy for Employment-Related Health Insurance Natl Tax J 2000; 53: 877 -887

11. **Cohen, SB.** Sample Design of the 1997 Medical Expenditure Panel Survey Household Component. AHRQ Pub. 1997-01.

12. **Cohen JW**. Design and Methods of the Medical Expenditure Panel Survey Household Component. AHCPR Pub. 1997:26.

13. **Machlin, SR and Taylor, AK**. Design, methods, and field results of the 1996 Medical Expenditure Panel Survey Medical Provider Component. AHRQ Pub. 2000: 28.

14. **Sommers J.** List Sample Design of the 1996 Medical Expenditure Panel Survey Insurance Component**.** AHCPR Pub. 1999:6.

16. **Moeller, J F. , Cohen SB, Mathiowetz, NA, Wun LM**. Regression-Based Sampling for Persons with High Health Expenditures: Evaluating Accuracy and Yield with the 1997 MEPS. 2003: *Medical Care*. Volume 4, No. 7: 44-52.

16 **Cox BG, Cohen SB.** Methodological issues for health care surveys. New York: Marcel Dekker; 1985.

17. **Adler, G**. A Profile of the Medicare Current Beneficiary Survey. Health Care Financing Review, 1994 Vol 15, No. 4: 153-163

18. **Cohen, S.B. and Carlson, B. L**. A Comparison of Household and Medical Provider Reported Expenditures in the NMES. 1994; the Journal of Official Statistics, Statistics Sweden. Vol. 10, No. 1, 3-29.

19. **Cohen, S.B**. Sample Design of the 1996 Medical Expenditure Panel Survey: Medical Provider Component. 1998: The Journal of Economic and Social Measurement. Vo1 26, 1-29.

20**. Wun, LM and Cohen SB**. An Estimation Strategy to Correct Household Reported Utilization Data for Reporting Errors Based on Medical Provider Data. 2000 Proceedings of the Health Policy Statistics Section, American Statistical Association.

Acknowledgements

The author wishes to thank Joel Cohen and William Yu for their careful review of the manuscript and helpful suggestions for additional content.