Enhancing the Medical Expenditure Panel Survey through Data Linkages

Lisa B. Mirel[1] and Steven R. Machlin[1]

[1]Agency for Healthcare Research and Quality,
540 Gaither Road, Rockville, MD 20850

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

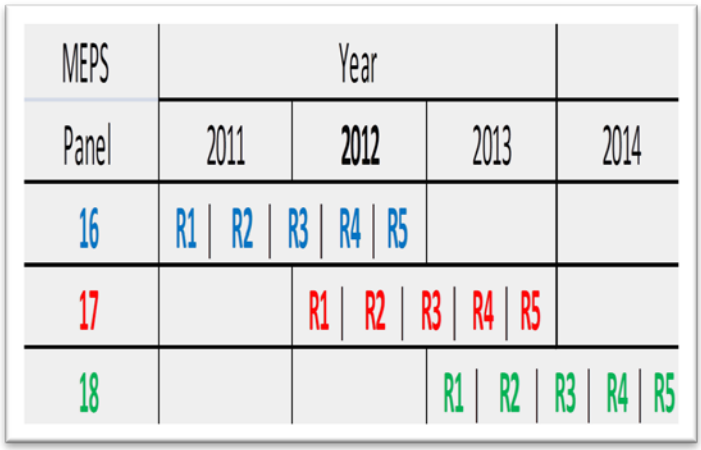
Introduction

Linkages between household surveys and other data sources can enhance the analytic capabilities of a survey. In particular, these linkages can reveal ways to improve data quality and can highlight areas for methodological studies. This paper will describe two data linkages to the Medical Expenditure Panel Survey Household Component (MEPS-HC). The first linkage stems from the sample design of the MEPS-HC. The MEPS-HC is a subsample of the National Health Interview Survey (NHIS), which creates a natural link between the two surveys. This linkage can be further extended to administrative records through the National Center for Health Statistics (NCHS) data linkage program. The second linkage deals with the subset of MEPS-HC sample persons who were selected and who signed permission forms to contact medical providers for the MEPS Medical Provider Component (MEPS-MPC). In order to combine these two MEPS components (MEPS-HC and MEPS-MPC) it is necessary to use a probabilistic linkage algorithm. We will discuss the details of the matching algorithm, including variables used in matching and the calculation of match weight scores, and highlight the implications of the linkage outcomes.

MEPS Household Component

The MEPS-HC is a complex, multi-stage, nationally representative sample of the U. S. civilian noninstitutionalized population. It has been an annual survey since 1996. Each year a new sample is drawn as a subsample of households that participated in the prior year's NHIS conducted by the NCHS, Centers for Disease Control and Prevention. MEPS-HC uses an overlapping panel design (Ezzati-Rice, Rohde, & Greenblatt, 2008). There are 5 rounds of data collection covering a two year reporting period. During each calendar year data are collected simultaneously for two MEPS panels. One panel is in its first year of data collection (e.g., in 2012, Rounds 1, 2, and 3 of Panel 17), and the prior year's panel is in its second year of data collection (e.g., in 2012, Rounds 3, 4, and 5 of Panel 16). The reference period for Round 3 for each MEPS-HC panel overlaps two calendar years. Annual estimates are made by combining data for the same calendar year from the panel in its first year of data collection and the panel in its second year of data collection (Figure 1).

Figure 1. MEPS-HC overlapping panel design, for Panels 16, 17, 18



Data are collected in the MEPS-HC through a series of five computer assisted personal interviewing (CAPI) systems on a variety of health related issues. These issues include health conditions, use of medical care services, charges and payments, and access to care. The MEPS-HC supports national annual estimates of health care use, expenditures, insurance coverage, sources of payment, access to care and health care quality. The MEPS-HC is a household level sample; data are collected for all target population members in the household.

I. MEPS-HC Integrated Design with NHIS

The integrated design of the MEPS-HC and NHIS allows for direct linkages between the two surveys. As background, the NHIS is a multi-purpose health survey that serves as the principal source of information on the health status and health behaviors of the civilian, noninstitutional U.S. population. NHIS uses a complex, multi-stage sample design, oversampling Asians, Hispanics and blacks (Division of Health Interview Statistics, 2011). This complex survey design carries over to the MEPS-HC through the set of NHIS responding households that comprise the frame for MEPS-HC sample selection.

The linkage between the MEPS-HC and NHIS is described below, focusing on four key areas of utility: the sample design, additional analytic variables, extension of the longitudinal data period, and enhancement with administrative record data.

Sample design. As noted above, the MEPS-HC is a nationally representative subsample of responding households from the previous year's NHIS. The integrated design permits the use of NHIS socio-demographic factors and health status variables to oversample certain subgroups in MEPS-HC. For example, NHIS race and ethnicity variables are used to stratify households in the MEPS-HC frame for oversampling of minorities. In 2011, an additional stratum based on responses to a cancer related question in the NHIS was incorporated into the MEPS-HC frame to enhance the sample size of cancer survivors for a MEPS cancer supplement. More specifically, the NHIS households that were eligible for the MEPS-HC were selected with certainty if they contained an NHIS respondent who had responded in the NHIS "sample adult" questionnaire that they had been diagnosed with some form of cancer. Research is also currently being done to assess the potential for using other NHIS variables to target Community Health Center Users for oversampling in future surveys.

Additional variables. The integrated design also enables the use of NHIS variables to further enhance the MEPS-HC data. For example, NHIS variables are used in the non-response adjustments for the statistical weights. A full list of these variables can be found in Methodology Report #24, "Estimation Procedures for the 2007 Medical expenditure Panel Survey Household Component" (Machlin, et al., September 2010). In addition, there are variables that are collected in NHIS and not in MEPS-HC (e.g. citizenship status and certain health conditions) that can be useful as additional correlates for analyses.

Extended longitudinal data period. The linkage between the two surveys can extend the longitudinal data period for certain analyses. One example of such an extended longitudinal period is contained in a statistical brief titled, "The Long-Term Uninsured in America, 2007-2010 (Selected Intervals): Estimates for the U.S. Civilian Noninstitutionalized Population under Age 65." While the estimates are derived based on the MEPS panel that covered 2009-2010, information was used from NHIS to further reflect health insurance status prior to 2009 (Rhoades & Cohen, January 2013). Some caution, however, must be taken when using the linked MEPS-HC and NHIS. Not all persons in the MEPS-HC final sample can be linked with the NHIS sample. This difference is primarily due to newly eligible persons joining the MEPS sample households after the NHIS interview (e.g. due to marriage, birth, return from the military, leave from an institution, or return to the U.S. after living overseas). Also, there is a gap in the extended longitudinal period covered by the linked data because the reference period for the MEPS-HC begins January 1st and completion of the NHIS occurs at some point in the prior year.

Administrative records. Previous research linked MEPS sample persons who received Medicare to their enrollment and claims data from the Centers for Medicare and Medicaid Services through special data use agreements (Zuvekas & Olin, 2009). These linkages were done using survey reported Medicare health insurance claim numbers or social security numbers. These unique identifiers are no longer collected in the MEPS-HC. However, the MEPS-HC and the NHIS integrated design enables the indirect linkage of some MEPS-HC sample persons to their administrative records through the NCHS data linkage program. NCHS has a record-linkage program designed to maximize the scientific value of the Center's population-based surveys. The NCHS program links various surveys (including NHIS) with administrative records. The administrative records include mortality

data from the National Death Index (NDI), Centers for Medicare and Medicaid Services enrollment and claims records, and Social Security Administration benefit history data. In one example, using the MEPS-HC linked to NHIS and administrative records, the association between medical expenditures reported in MEPS-HC in 2001 and mortality in the interval 2002-2006 based on the NHIS linked to the NDI was assessed (Cohen, 2012). It showed that medical expenses in 2001 were higher for those who died in the interval 2002-2006 compared to those assumed alive at the end of 2006. Another example of linked MEPS-HC and the NHIS and administrative data involves a research project that will be conducted at the NCHS Research Data Center in 2014. The goal of the project is to obtain more up-to-date, payment-to-charge ratios that will improve the expenditure imputations in MEPS for Medicaid recipients for certain types of services. These analyses would be much more challenging to conduct without the integrated design and the data linkage to administrative records.

II. Linkage Between MEPS-HC and MEPS Medical Provider Component

In this section we shift attention away from the integrated design with NHIS to the within MEPS linkage between the MEPS Household (MEPS-HC) and Medical Provider (MEPS-MPC) Components, with particular focus on the matching algorithm used to link the two components. The MEPS-MPC is a sample of medical providers reported in the MEPS-HC as providing care to sample persons. For a provider or pharmacy to be contacted for the MPC, Health Insurance Portability and Accountability Act (HIPAA) compliant permission forms must be signed by the adult family member who received services (or by the parent for a child). In general, the MPC collects data of higher quality that are difficult for household respondents to report completely or accurately, such as charges and payments, dates of visits, diagnosis and procedure codes. The primary uses of the MPC are to supplement or replace expenditure data reported in the MEPS-HC, to serve as an imputation source, and to support methodological research.

The MPC sample is based on health care events reported by household respondents. The sample includes all hospitals (and associated doctors) reported as providing inpatient, outpatient or emergency care, all home health agencies, all pharmacies, and about 50% of the office based physicians reported as providing care/medicines to sample persons in the MEPS-HC. The providers are contacted and asked about all events for the sample person during the calendar year. The MEPS-MPC is not designed as an independent nationally representative sample of medical providers.

Matching algorithm. Unique identifiers do not exist for medical events that would allow a direct linkage between the MEPS-HC and MPC. For example, due to possible respondent recall error, the date of a medical event reported by the household respondent may differ from the date reported by the medical provider. Therefore, the linkage between the two files uses a probabilistic linkage based on the Fellegi and Sunter algorithm (Fellegi & Sunter, 1969).

Matching is done in two passes within block groups for all possible event pairs. A block group is a non-overlapping group used in matching to reduce the number of all possible pair comparisons. The first pass of matching defines the block group as the person and contact group. A contact group is a cluster of providers sharing the same point of contact. Pass 1 matching exhausts all possible matches in the block group and selects the best set of matched event pairs. Then, only the events not matched in Pass 1 are eligible for Pass 2. The block group in Pass 2 is the person; this allows for events to be matched across provider IDs to allow for potential error across IDs (which may not have been grouped together for Pass 1) or potential misreporting by the household as to which provider they saw.

The algorithm links records based on comparison scores that are computed from each matching variable. For hospital and office based physician events the variables being compared are:

- date of the event,
- type of event (inpatient, out-patient, emergency room, and office based),
- flags indicating whether surgery, radiology and laboratory procedures were performed,
- length of the hospital stay (for hospital inpatient events),
- indicator if the event was part of a global fee, and
- condition codes (e.g. headache, cancer, infectious disease, etc.).

Each match variable is assigned either an agreement score if the HC and MPC agree (positive in value) or a disagreement score if they disagree (negative in value) for each event pair. The agreement and disagreement scores are functions of conditional probabilities estimated for each match variable. These probabilities, "m" and "u," are defined as follows:

m_{ijk}=Pr{field i agrees in the event pair (j,k), given the pair is in the set of M true matches}
u_{ijk}=Pr{field i agrees in the event pair (j,k), given the pair is in the set of U true non-matches}.

Therefore, "m_{ijk}" is the probability that field *i* agrees among the truly matched records and "u_{ijk}" is the probability that field *i* agrees at random. The m_{ijk} and u_{ijk} are estimated from an iterative algorithm. The initial value of m_{ijk} is set to a specified constant, 0.90. The initial value of u_{ijk} is initially set at:

u_{ijk}=# times a field agrees/Total number of pairs possible.

The algorithm continues to iterate until successive differences in the m_{ijk}'s and u_{ijk}'s are sufficiently small (0.001).

The agreement field scores are calculated for (i,j,k) as base-2 logarithm of the ratio of these two probabilities:

$$A_{ijk}=\log_2 [m_{ijk}/u_{ijk}]$$

And the disagreement field scores are calculated for (i,j,k) as base-2 logarithm of the ratio of one minus the probability of each:

$$D_{ijk}=\log_2 [(1-m_{ijk})/(1-u_{ijk})]$$

The score is then calculated as:

$$S_{ijk}=\begin{cases} A_{ijk} \text{ if field i agrees} \\ D_{ijk} \text{ if field i disagrees} \end{cases}$$

Once all of the comparison scores are computed, a match weight score for the event pair is formed by summing them, as noted below:

$$W_{jk}=\sum_{i} S_{ijk}$$

As the sum of the comparison scores, the match weight score quantifies the strength (i.e., quality) of the MEPS-HC and MPC event pair, because the comparison scores are based on the likelihood that the MEPS-HC information and MPC information refer to the same event. The match weight score is calculated for all possible event pairs. The best set, that is, the set of event pairs giving the maximum sum of event-pair match weights, is selected by the Hungarian method (aka Kuhn-Munkers algorithm). The Hungarian Method is a solution to what is sometimes referred to as the linear assignment problem or the linear sum assignment problem (Goldberger & Tassa, 2008). Identifying the maximum sum of weights is subject to certain constraints; particularly, an event can only be included in one pair.

After the matching process is complete MEPS-HC incorporates the expenditure information obtained from linking to the MPC to the maximum extent possible. Since MEPS-HC expenditure data is generally less accurate or complete than that obtained in the MEPS-MPC, maximizing use of MEPS-MPC information should generally improve the quality of the overall MEPS expenditure estimates.

Methodological studies. The linkage of the MEPS-HC and MPC has led to improved survey expenditure estimates, and also to a variety of methodological research that can help inform policy decisions and survey operations. Previous research using this type of linked data assessed trends in provider capitation by examining differences in capitation by socio-demographic characteristics including age, race and ethnicity, sex, health status, family income, census region, and urbanicity (Zuvekas & Cohen, 2010). In addition, previous work analyzed provider reported data to create adjustments for comparing health expenditure estimates from the 1996 MEPS to the 1987 NMES

(Zuvekas & Cohen, 2002). There has also been research comparing household-reported data on expenditures to expenditures reported in the MPC using the 1987 National Medical Expenditure Survey (Cohen & Lepidus Carlson, 1994) and early years of MEPS (Machlin, Cohen, Zuvekas, & Thorpe, 1999). Similar analyses are currently being conducted using data from the 2010 and 2011 MEPS-HC and MPC.

Summary

In summary, this paper highlights two types of data linkages to the MEPS-HC that enhance the survey. The first, the integrated sample design of the MEPS-HC and NHIS enables data linkages between the two surveys for enhanced analyses which can be further extended to administrative records through the NCHS data linkage program. The second, the linkage between MEPS-HC and the MPC, improves the quality of MEPS expenditure data and provides valuable data for methodological studies.

References

Cohen, S. B. (2012). The Utility of the Integrated Design of the Medical Expenditure Panel Survey to Inform Mortality Related Studies. *In JSM Proceedings, Section of Survey Research Methods* (pp. 3461-3470). Alexandria, VA: American Statitsical Association.

Cohen, S. B., & Lepidus Carlson, B. (1994). A Comparison of Household and Medical Provider Reported Expenditures in the 1987 NMES. *Journal of Official Statistics, 10*(1), 3-29.

Division of Health Interview Statistics. (2011, June 15). *2010 NHIS Survey Description Document.* Retrieved December 13, 2013, from NCHS CDC: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2010/srvydesc.pdf

Ezzati-Rice, T. M., Rohde, F., & Greenblatt, J. (2008). *Sample Design of the Medical Expenditure Panel Survey Household Component, 1998–2007.* Agency for Healthcare Research and Quality. Rockville, MD: Methodology Report No. 22.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*, 1183-1210.

Goldberger, J., & Tassa, T. (2008). A hierarchical clustering algorithm based on the Hungarian methods. *Pattern Recognition Letters, 29*, 1632-1638.

Machlin, S., Chowdhury, S., Ezzati-Rice, T., DiGaetano, R., Gosksel, H., Wun, L.-M., et al. (September 2010). *Estimation Procedures for the Medical Expenditure Panel Survey Household Component. Methodology Report #24.* Rockville, MD: Agency for Healthcare Research and Quality.

Machlin, S., Cohen, J., Zuvekas, S., & Thorpe, J. (1999). Accuracy of Household Reported Payments for Physician Visits in the 1996 Medical Expenditure Panel Survey. *American Statistical Association Proceedings.*

Rhoades, J. A., & Cohen, S. B. (January 2013). *The Long-Term Uninsured in America, 2007-2010 (Selected Intervals): Estimates for the U.S. Civilian Noninstitutionalized Population under Age 65. Statistical Brief #399.* Rockville, MD: Agency for Healthcare Research and Quality.

Zuvekas, S. H., & Cohen, J. W. (2002). A Guide to Comparing Health Care Expenditures in the 1996 MEPS to the 1987 NMES. *Inquiry, 39*(1), 76-86.

Zuvekas, S. H., & Cohen, J. W. (2010). Paying Physicians By Capitation: Is The Past Now Prologue? *Health Affairs, 29*(9), 1661-1666.

Zuvekas, S. H., & Olin, G. L. (2009). Accuracy of Medicare Expenditures in the Medical Expenditure Panel Survey. *Inquiry, 46*(1), 92-108.

Zuvekas, S. H., & Olin, G. L. (2009). Validating Household Reports of Health Care Use in the Medical Expenditure Panel Survey. *Health Services Research, 44*(5 Part I), 1679-1700.