Use of Paradata to Evaluate Medical Expenditure Panel Survey Data and Operations

Lisa B. Mirel[1], Steven R. Machlin[1]
[1]Division of Statistical Research and Methods, Center for Financing, Access and Cost Trends, Agency for Healthcare Quality and Research, 540 Gaither Road, Rockville, MD 20852

Abstract

Paradata in survey research have become increasingly used in recent years to facilitate the monitoring of survey operations and to improve data quality. Paradata consist of information about the data collection process in a survey, including interviewer observations, interview language, computer generated time variables for questionnaire sections and numerous other variables. One survey that uses paradata to monitor survey operations and explore improvements in data quality is the Medical Expenditure Panel Survey Household Component (MEPS-HC). The MEPS-HC is a complex, multi-stage, nationally representative sample of the U.S. civilian non-institutionalized population with an overlapping panel design. Each year a new sample is drawn as a subsample of households that participated in the prior year's National Health Interview Survey (NHIS) (conducted by the National Center for Health Statistics). Data are collected in the MEPS-HC through a series of five computer assisted personal interviews (CAPI) that cumulatively cover a two year period on a variety of health related issues, including health conditions, use of medical care services, charges and payments, and access to care. There is a wealth of MEPS-HC paradata associated with the multiple MEPS-HC interviews and additional paradata information can be obtained by linking to the prior year's NHIS. This paper describes an evaluation of the association between paradata measures and reporting of health care events in the MEPS-HC, with particular focus on two paradata variables (CAPI data collection path and memory aid and record usage) and their association with reported office-based health care utilization. The results are interpreted in the context of using paradata to identify areas that may be at higher risk for under-reporting of health care utilization.

Key Words: Medical Expenditure Panel Survey (MEPS), Paradata, Health Care Utilization

1. Introduction[1]

Paradata in survey research have become increasingly used in recent years to facilitate monitoring of survey operations and improve data quality. Paradata consist of information about the data collection process in a survey, including interviewer observations, interview language, computer generated time variables for questionnaire sections and numerous other variables. They reflect processes that the survey administrators can control for greater uniformity, if needed (Nicolaas, 2011). Results

[1] The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality are intended or should be inferred.

from paradata analyses can help to assess sources of survey error like non-response, measurement and processing errors which can inform improvements to survey estimates.

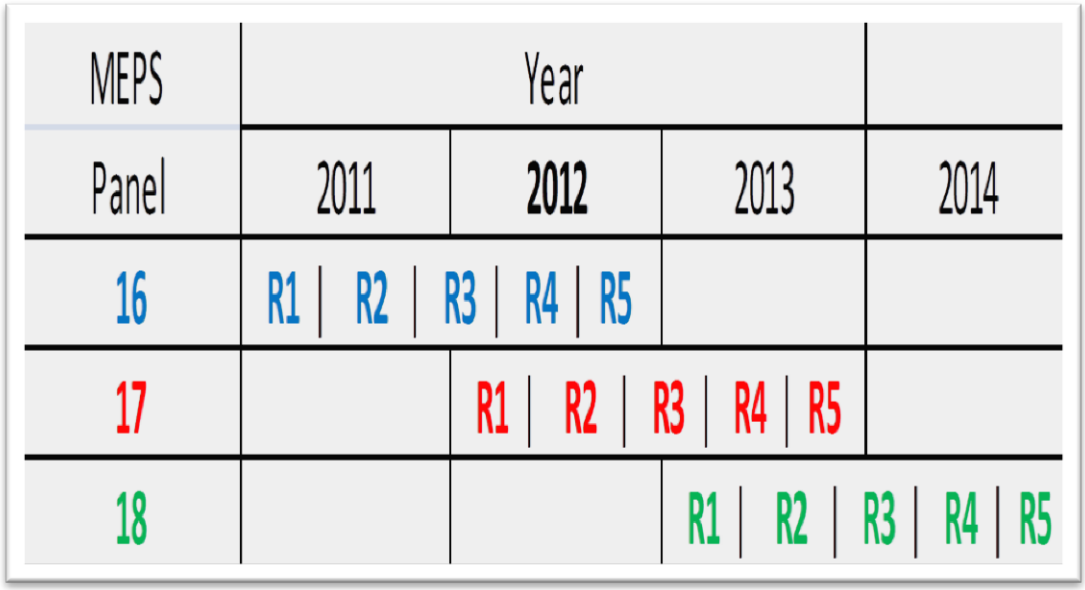
One survey that uses paradata to monitor survey operations and explore improvements in data quality is the Medical Expenditure Panel Survey Household Component (MEPS-HC). This paper, using data from the first interview of the MEPS panel that was initiated in 2012 (Panel 17 Round 1), describes an evaluation of the association between paradata measures and reporting of health care events. Previous research indicates there is under-reporting of health care events in household surveys (Zuvekas & Olin, 2009). The goal of this analysis is to utilize the paradata to identify data collection methods that are at higher risk for under-reporting. In this family level analysis we focus on two specific paradata variables- CAPI data collection path and memory aid and record usage- and their associations with reporting of office based medical events. The two main paradata variables are related in that the use of records is a component of each variable. However, they are collected independently in the instrument and are not necessarily consistent. Although survey administrators cannot directly control how the interviewers enter data in the CAPI or what records the respondents use, there are training initiatives that can be implemented to create more uniformity among the interviewers and data collection.

1.1 Background

The MEPS-HC is a complex, multi-stage, nationally representative sample of the U.S. civilian non-institutionalized 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 National Health Interview Survey (NHIS) (conducted by the National Center for Health Statistics). 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.

MEPS-HC uses an overlapping panel design (Ezzati-Rice, Rohde, & Greenblatt, 2008). During each calendar year data are collected simultaneously for two MEPS-HC panels. One panel is in its first year of data collection (e.g., in 2012, Rounds 1, 2, and 3 of Panel 17), while 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 interviews on a variety of health related issues, including health conditions, use of medical care services, charges and payments, and access to care.

The survey is typically completed by one respondent for all members of the family. A family in MEPS-HC is defined as two or more persons living together in the same household who are related by blood, marriage, adoption, foster care or have identified as a single unit. The respondent is asked to report about all health care use for the family members during the reference period.

In the MEPS-HC paradata are collected at the family level as the survey data are being collected. There is a wealth of MEPS-HC paradata associated with the multiple MEPS-HC interviews, and additional paradata and other information can be obtained by linking to the prior year's NHIS. Some of the paradata variables collected in MEPS-HC include the number of contacts, the language of the interview, and whether the interview was collected in person or by phone. Memory aid and record usage and data collection path are also paradata that are ascertained.

2. Methods

This analysis is based on Panel 17 Round 1 paradata for MEPS-HC families that linked to the 2012 Point in Time (PIT) file (http://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-143) (Note: the paradata variables are not on the public use file and therefore need to be linked by family ID to the PIT file.) In addition, we only included those observations with non-missing values for the variables used in the analysis.

The paradata variables are collected at the family level and offer the first glimpse of reported health care utilization for that calendar year. We constructed an outcome variable to be an annualized estimate of the mean number of office based events per

person (Figure 2). Office based events were used for the analysis since they are the most commonly reported event type in the MEPS-HC.

Figure 2. Calculation of the dependent variable.

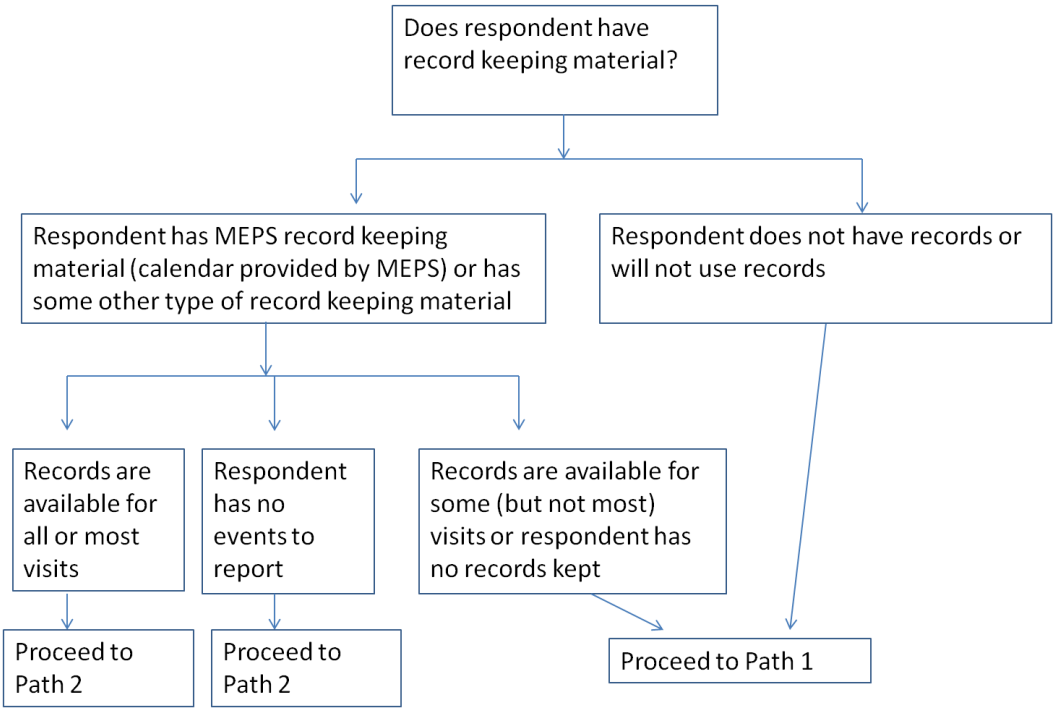
$$Y=\left[\frac{\#\ events}{\#\ people\ in\ family}\right] \Big/ \#\ days\ in\ reference\ period \Big] * 365$$

In order to assess the association of the CAPI data collection path and memory aid and record usage on the reporting of office based health care events, we began by looking at descriptive statistics. In addition to these two variables, we assessed other variables that are commonly associated with health care event reporting. Results are presented, using descriptive statistics, for paradata and non-paradata variables. The unadjusted annual mean office based visits per person and the percent distribution of each variable is presented. In addition, a multivariable regression model was used to assess if either of the main analytic paradata variables of interest- CAPI data collection path and memory aid and record usage- were associated with office based utilization after adjusting for other covariates associated with utilization. All covariates used in the analyses are enumerated in Sections 2.1 and 2.2 below. Beta coefficients, adjusted means and their standard errors are presented for the main paradata variables of interest. All analyses used the family level weight based on the 2012 PIT file and were run in SUDAAN to account for the complex design of the MEPS-HC. Adjusted means were calculated using the least squares means option in SUDAAN's Proc Regress, and comparisons were made for the different categories of a variable compared to a referent group. Statistical significance was set at a p-value < 0.05.

2.1 Main paradata variables

The *data collection path* indicates which of the two CAPI paths were used by the interviewer to enter reported healthcare events. The path is determined by the interviewer based on the record keeping materials provided by the respondent and questioning to determine completeness of the records. If the respondent does not have any records then the interview proceeds to a more intensive probing data collection path (path 1). This path involves more extensive person by person probes and questions than the alternative path (path 2) and, therefore takes longer. In path 2 events are entered directly into the CAPI based on the records provided by the respondent. This path is generally less burdensome for the interviewer and respondent. There is some probing for more events but not as extensively as in path 1. The variable used in this analysis to indicate data collection path was created based on answers to questions in the calendar path section of the questionnaire as illustrated in Figure 3 below. (http://meps.ahrq.gov/mepsweb/survey_comp/hc_survey/2011/CA110311.pdf). In order to assess the utilization reporting based on the data collection path a three level variable was constructed, those with records in path 2, those who said they had no events to report prior to entering path 2, and those in path 1.

Figure 3. Determination of data collection path in the MEPS-HC



Memory aid and record usage is intended to increase the completeness of event reporting. Respondents are encouraged to use a calendar, explanation of benefits (EOB) from their insurance company, other insurance records or any other aid that may help them remember health care events for themselves and family members. The MEPS memory aid and record use paradata variable was created based on a series of questions that the interviewer completed at the end of the closing section in the interview. The question asks, "Were any of the following memory aids used by the respondent during the interview?" and proceeds to list a variety of memory aids and records, including but not limited to a calendar (prior to the start of the interview families are given a calendar to record health care events of all family members), electronic records, and insurance payment statement/ explanation of benefits (EOB). The interviewer can check all that apply. A complete list of memory aids and records can be found in the closing section questionnaire (http://meps.ahrq.gov/mepsweb/survey_comp/hc_survey/2011/CL110311.pdf). For this analysis a four level summary variable comprised of the following categories was created: 1) calendar only, 2) calendar and other records, 3) other records only (not calendar) and 4) no records. Memory aids and records for prescription drugs (pharmacy patient profile and medicine bottle/ receipt) were not included as part of the memory aid and record use variable created for this analysis since our analysis only focused on office based events and not prescribed medicine purchases.

2.2 Other variables

Other paradata variables used in the analysis include:

- language of the interview,
- whether the interview was collected in person or by phone (only about 6% of families had phone interviews),

- whether the interview was conducted in a multi or single session,
- if the case was transferred to another interviewer because the family moved or for some other reason,
- number of contacts,
- experience level of the interviewer,
- respondents' reluctance to respond (a respondent is classified as reluctant if at any time anyone in the family refused to participate),
- supervisor ID,
- reference period length, and
- completeness of the NHIS interview (classified as partial or completed interview).

In addition, the non-paradata variables included:

- family income as reported in NHIS,
- core based statistical area (CBSA)/metropolitan statistical area (MSA) status based on NHIS,
- reported race/ethnicity of the family members,
- number of people in the family,
- percentage of people in the family 0-15 years and over 64 years,
- marital status,
- insurance status of the family members, and
- number of priority health conditions.

The table in the Appendix includes the descriptive statistics of these variables used in the analysis and their corresponding categories.

3. Results

3.1 Descriptive Statistics

Based on the MEPS paradata file linked to the 2012 PIT file, in Panel 17 Round 1, there were 7,510 families with non-missing data for all of the variables used in the analysis. The average number of people per family was about 2. The overall annualized mean number of office based events per person was 6.4 (standard error=0.21).

The percent distributions, estimated annual mean number of office based events per person and standard errors are provided for the main paradata variables of interest in Table 1 below. Less than half (44.5%) of families did not use any memory aids or records during the interview and 42.4% were interviewed via the path 1. Unadjusted annualized mean office visits per person were lowest for interviews where no memory aids or records were used (3.3) and for those in path 2 with no events to report (1.0). However, it should be noted that even though the respondent responded no events to report prior to entering the path 2, the average number of events per year was 1, suggesting the minimal probes in this path may have increased reporting of health care utilization.

Table 1. Percent distribution, estimated annual mean number of office based visits per person, and standard errors for main paradata variables of interest, MEPS paradata file Panel 17, Round 1.

Covariate	Percent Distribution (SE)	Annual mean office based visits per person (SE)
Data collection path		
Path 2- all/most records	44.8 (1.11)	9.4 (0.37)
Path 2- no events to report	12.8 (0.67)	1.0 (0.22)
Path 1	42.4 (1.25)	4.7 (0.28)
Memory aid and record use		
Calendar only	23.2 (0.97)	8.9 (0.51)
Calendar and other records	20.4 (1.11)	10.3 (0.57)
Other records (not calendar)	11.9 (0.61)	5.9 (0.55)
No records	44.5 (1.23)	3.3 (0.20)

Note that the same types of estimates are provided for other paradata and non-paradata variables in the table in the Appendix.

3.2 Results from multivariable model

Table 2 presents the beta coefficients, adjusted means and standard errors for the two paradata variables of interest. Interviews that used path 2 and had all or most records had 2.6 visits more per year on average than those that used path 1(8.4 versus 5.8). Respondents that initially stated there were no events to report and entered path 2 had an estimated average of 1.6 visits per year. The adjusted averages of annualized visits per person for those who used a calendar and other records, a calendar only or records other than a calendar, were significantly higher than the average for those who did not use any records (4.3).

Table 2. Beta coefficients, standard errors and estimated adjusted mean of annualized office based visits per person for paradata variables of interest, MEPS paradata file Panel 17, Round 1.

Covariate	Beta coefficient (SE)*	Adjusted mean of annualized office based visits per person*
Data collection path		
Path 2- all/most records	2.8 (0.48)^	8.4 (0.33)^
Path 2- no events to report	-4.0 (0.37)^	1.6 (0.29)^
Path 1 (referent group)	Ref	5.6 (0.29)
Memory aid and record use		
Calendar only	3.8 (0.55)^	8.1 (0.48)^
Calendar and other records	4.8 (0.61)^	9.1 (0.53)^
Other records (not calendar)	1.3 (0.56)^	5.7 (0.54)^
No records (referent group)	Ref	4.3 (0.24)

*Multivariable model included interview language, supervisor ID, mode of interview, multi/single session interview, case transfer, interviewer experience, reference period length, NHIS income, NHIS completion status, CBSA/MSA status from NHIS, race/ethnicity, percent 0-15 years, percent over 64 years, marital status, health insurance

status, reluctant respondent, number of contacts, number of people in the family, and priority conditions.

^ p-value<0.05 compared to referent group

4. Discussion

There is a wealth of paradata variables in the MEPS-HC. In this analysis we focused on two key variables which could potentially be used to identify areas where data collection could be improved. Results from our model should be interpreted within data limitations. The model was conducted at the family level (where paradata variables are collected) and not at the person level where health care utilization is typically ascertained. Nevertheless, the results shed light on important issues with data collection. Holding all other variables constant in a multivariable model, we found that going down the path 1 and not using memory aids or records resulted in lower reporting of office based medical events. These results may be partially attributable to families with no utilization being more likely to be interviewed in path 1 and/or to legitimately having no records. We noted that for those who initially reported no events to report and continued down path 2 (about 13 percent of families) had 1.6 events on average. This suggests that even the less intensive probing in this path may have improved their utilization reporting. This also suggests that their reporting may have increased even more if they had gone down the path 1 and received person by person probes. The results from this analysis have helped to inform interviewer training initiatives to identify areas that may need improvement with respect to under-reporting of health care utilization. In recent months, there has been a great deal of outreach to the interviewers working on the MEPS-HC, including home study programs and refresher trainings. Two goals of the trainings are to 1) raise awareness among interviewers for how to choose the best data collection path given the record keeping materials provided by the respondent at the start of the interview and 2) work with interviewers on new strategies to encourage respondents to maintain and use records. These training efforts are on-going and continue to be monitored using paradata.

References

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.

Nicolaas, G. (2011). *Survey Paradata: A Review.* Swindon, England: Economic and Social Research Council.

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.

Appendix

Percent distribution, estimated annual mean number of office based visits per person, and standard errors for other variables in multivariable model, MEPS paradata file Panel 17, Round 1.

Covariate	Percent Distribution (SE)	Annual mean office based visits per person (SE)
OTHER PARADATA VARIABLES		
Language		
Spanish	5.3 (0.40)	2.7 (0.19)
Both English/ Spanish	1.1 (0.15)	6.4 (1.78)
Other language	0.6 (0.16)	4.7 (1.46)
English	93.0 (0.52)	6.6 (0.22)
Mode of Interview		
Telephone	5.7 (0.39)	3.0 (0.47)
In person	94.3 (0.39)	6.6 (0.22)
Multi/single session interview		
Multi	9.9 (0.67)	7.7 (0.68)
Single	90.1 (0.67)	6.2 (0.21)
Transfer		
Case was not transferred	74.5 (0.99)	6.8 (0.25)
Case was transferred	25.5 (0.99)	5.0 (0.36)
Number of contacts		
>=4	61.0 (1.02)	59.7
<4	39.0 (1.02)	40.3
Interviewer experience level		
Experienced	86.9 (1.27)	6.5 (0.22)
New	13.1 (1.27)	5.4 (0.46)
Reluctant Respondent		
Yes	11.4 (0.53)	4.4 (0.49)
No	88.6 (0.53)	6.6 (0.22)
NHIS Outcome Summary		
Partial	17.8 (0.78)	4.9 (0.45)
Complete	82.2 (0.78)	6.7 (0.41)
Reference period length (months)		
0-<3	69.9 (0.78)	7.2 (0.27)
3-<4	12.9 (0.60)	5.3 (0.49)
4-<5	7.7 (0.51)	4.2 (0.47)
>=5	9.5 (0.48)	3.5 (0.33)

NON-PARADATA VARIABLES		
NHIS income category		
Missing	20.2 (0.70)	5.9 (0.47)
0-<35,000	32.7 (0.96)	6.2 (0.34)
35,000-<50,000	11.0 (0.48)	6.5 (0.73)
50,000-<75,000	14.2 (0.57)	6.7 (0.42)
>=75,000	22.0 (0.70)	6.7 (0.41)
CBSA/MSA status in NHIS		
Principal city of the CBSA/MSA	34.8 (1.35)	6.3 (0.33)
In CBSA/MSA but not principal city	48.8 (1.50)	6.7 (0.31)
Not in CBSA/MSA	16.4 (1.40)	5.6 (0.41)
Race/ethnicity of family members		
Asian among race/ethnicity reported	5.1 (0.40)	4.0 (0.54)
Hispanic among race/ethnicity reported	14.7 (0.81)	4.1 (0.21)
Black among race/ethnicity reported	12.3 (0.66)	5.2 (0.34)
White/other among race/ethnicity reported	67.9 (0.95)	7.2 (0.28)
Age of family members		
>=50% of family members 0-15 years	16.1 (0.53)	4.5 (0.31)
<50% of family members 0-15 years	83.9 (0.53)	6.7 (0.24)
>=50% of family members over 64 years	21.6 (0.77)	11.2 (0.71)
<50% of family members over 64 years	78.4 (0.77)	5.0 (0.16)
Number of people in the family		
>=4	20.5 (0.64)	3.7 (0.20)
<4	79.5 (0.64)	7.0 (0.25)
Marital status		
Family contains a married couple	47.6 (0.90)	6.2 (0.31)
Family does not contain a married couple	52.4 (0.90)	6.5 (0.27)
Health Insurance		
All family members covered	69.9 (0.78)	7.7 (0.28)
One or more family members not covered	30.1 (0.78)	3.2 (0.18)
None of the family members covered	12.8 (0.49)	2.6 (0.27)
At least one family member covered	87.2 (0.49)	6.9 (0.23)
Number of priority conditions		
0	22.1 (0.73)	2.9 (0.27)
1 or 2	38.6 (0.72)	5.3 (0.26)
>=3	39.4 (0.80)	9.3 (0.41)

Note: this table does not include supervisor identification number due to the large number of categories