MEPS HC-036:
1996-2020 Pooled Linkage Variance Estimation File

August 2022

Agency for Healthcare Research and Quality
Center for Financing, Access, and Cost Trends
5600 Fishers Lane
Rockville, MD 20857
(301) 427-1406

Table of Contents

A. Data Use Agreement

Individual identifiers have been removed from the micro-data contained in these files. Nevertheless, under sections 308 (d) and 903 (c) of the Public Health Service Act (42 U.S.C. 242m and 42 U.S.C. 299 a-1), data collected by the Agency for Healthcare Research and Quality (AHRQ) and/or the National Center for Health Statistics (NCHS) may not be used for any purpose other than for the purpose for which they were supplied; any effort to determine the identity of any reported cases is prohibited by law.

Therefore, in accordance with the above referenced Federal Statute, it is understood that:

1. No one is to use the data in this data set in any way except for statistical reporting and analysis; and

2. If the identity of any person or establishment should be discovered inadvertently, then (a) no use will be made of this knowledge, (b) the Director Office of Management AHRQ will be advised of this incident, (c) the information that would identify any individual or establishment will be safeguarded or destroyed, as requested by AHRQ, and (d) no one else will be informed of the discovered identity; and

3. No one will attempt to link this data set with individually identifiable records from any data sets other than the Medical Expenditure Panel Survey or the National Health Interview Survey.

By using these data you signify your agreement to comply with the above stated statutorily based requirements with the knowledge that deliberately making a false statement in any matter within the jurisdiction of any department or agency of the Federal Government violates Title 18 part 1 Chapter 47 Section 1001 and is punishable by a fine of up to $10,000 or up to 5 years in prison.

The Agency for Healthcare Research and Quality requests that users cite AHRQ and the Medical Expenditure Panel Survey as the data source in any publications or research based upon these data.

B. Background

1.0 Household Component

The Medical Expenditure Panel Survey (MEPS) provides nationally representative estimates of health care use, expenditures, sources of payment, and health insurance coverage for the U.S. civilian non-institutionalized population. The MEPS Household Component (HC) also provides estimates of respondents' health status, demographic and socio-economic characteristics, employment, access to care, and satisfaction with health care. Estimates can be produced for individuals, families, and selected population subgroups. The panel design of the survey, which usually includes 5 Rounds of interviews covering 2 full calendar years, provides data for examining person level changes in selected variables such as expenditures, health insurance coverage, and health status. Note that due to the COVID-19 pandemic, 2020 data collection moved primarily to phone rather than in-person. This posed a challenge in Panel 25 Round 1, which was difficult to start via phone as all phone numbers were not available for a new panel, resulting in a low response rate. To balance this and increase the number of completes to be comparable to previous years, Panel 23 was extended another year to 2020 with seven rounds of data collection. Using computer assisted personal interviewing (CAPI) technology, information about each household member is collected, and the survey builds on this information from interview to interview. All data for a sampled household are reported by a single household respondent.

The MEPS-HC was initiated in 1996. Each year a new panel of sample households is selected. Because the data collected are comparable to those from earlier medical expenditure surveys conducted in 1977 and 1987, it is possible to analyze long-term trends. Each annual MEPS-HC sample size is about 15,000 households. Data can be analyzed at either the person or event level. Data must be weighted to produce national estimates.

The set of households selected for each panel of the MEPS HC is a subsample of households participating in the previous year's National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics. The NHIS sampling frame provides a nationally representative sample of the U.S. civilian noninstitutionalized population and reflects an oversample of blacks and Hispanics. In 2006, the NHIS implemented a new sample design, which included Asian persons in addition to households with black and Hispanic persons in the oversampling of minority populations. MEPS further oversamples additional policy relevant sub-groups such as low income households. The linkage of the MEPS to the previous year’s NHIS provides additional data for longitudinal analytic purposes.

2.0 Medical Provider Component

Upon completion of the household CAPI interview and obtaining permission from the household survey respondents, a sample of medical providers are contacted by telephone to obtain information that household respondents cannot accurately provide. This part of the MEPS is called the Medical Provider Component (MPC) and information is collected on dates of visit, diagnosis and procedure codes, charges and payments. The Pharmacy Component (PC), a subcomponent of the MPC, does not collect charges or diagnosis and procedure codes but does

collect drug detail information, including National Drug Code (NDC) and medicine name, as well as date filled and sources and amounts of payment. The MPC is not designed to yield national estimates. It is primarily used as an imputation source to supplement/replace household reported expenditure information.

3.0 Survey Management

MEPS HC and MPC data are collected under the authority of the Public Health Service Act. Data are collected under contract with Westat, Inc. (MEPS HC) and Research Triangle Institute (MEPS MPC). Data sets and summary statistics are edited and published in accordance with the confidentiality provisions of the Public Health Service Act and the Privacy Act. The National Center for Health statistics (NCHS) provides consultation and technical assistance.

As soon as data collection and editing are completed, the MEPS survey data are released to the public in staged releases of summary reports, micro data files, and tables via the MEPS web site: https://meps.ahrq.gov/mepsweb/. Selected data can be analyzed through MEPSnet, an on-line interactive tool designed to give data users the capability to statistically analyze MEPS data in a menu-driven environment.

Additional information on MEPS is available from the MEPS project manager or the MEPS public use data manager at the Center for Financing Access and Cost Trends, Agency for Healthcare Research and Quality, 5600 Fishers Ln, Rockville, MD 20857 (Ph: 301-427-1406).

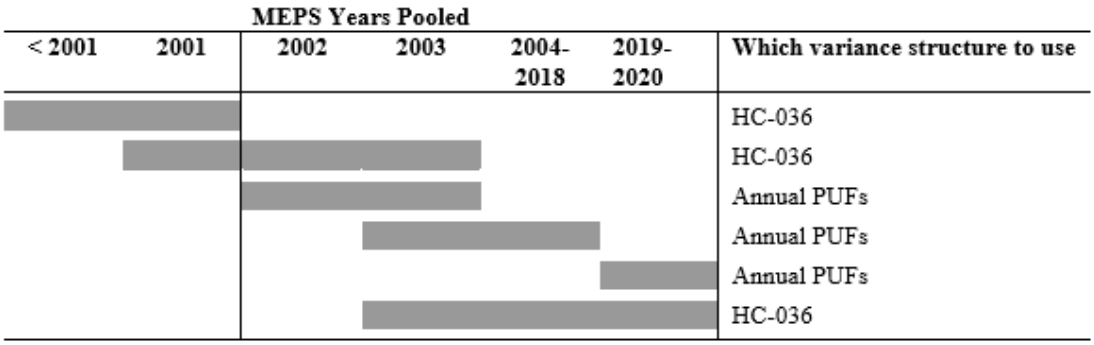
C. Technical and Programming Information

1.0 General Information

To facilitate analysis of subpopulations and/or low prevalence events, it may be desirable to pool together (i.e. combine) more than one year of MEPS-HC data to yield sample sizes large enough to generate reliable estimates. MEPS-HC samples in most years are not completely independent because households are drawn from the same sample geographic areas and many persons are in the sample for two consecutive years (see MEPS-HC Methodology Reports for more details at https://meps.ahrq.gov/mepsweb/). Despite this lack of independence, it is valid to pool multiple years of MEPS-HC data and keep all observations in the analysis because each year of the MEPS-HC is designed to be nationally representative. However, to obtain appropriate standard errors when pooling years of MEPS-HC data, it is necessary to specify a common variance structure that properly reflects the complex sample design of the MEPS.

This HC-036 file contains the proper variance structure to use when making estimates from MEPS data that have been pooled over multiple years and where one or more years are from 1996-2001 or 2019-2020. Prior to 2002, each annual MEPS public use file was released with a variance structure unique to the particular MEPS sample in that year. Also, the yeas 2019 and 2020 have a common but slightly different variance structure than other years. The variance structure in this HC-036 file reconciles the differences in the variance units between the units on the released annual MEPS public use files.

Between 2002 and 2018, the annual MEPS public use files were released with a common variance structure that allows users to pool data from 2002 to 2018. Also, between 2019 and 2020, the MEPS public use files were released with another common variance structure that allows users to pool data from 2019 and 2020. However, these common variance structures are not compatible with each other nor the structure on the annual PUFs released prior to 2002. Therefore, it is necessary to use the variance structure on this HC-036 dataset when pooling data from MEPS years prior to 2002 and also from 2019 - 2020. The following image presents some scenarios to clarify when analysts should use the variance structure in this HC-036 file.



In the first scenario, only MEPS data from years prior to 2002 are pooled together. In this case, analysts must use the variance structure in HC-036. In the second scenario, data from years prior to 2002 are pooled together with data from 2002 and forward up to 2018. The variance structure from HC-036 must be used in this circumstance as well. In the third and fourth scenarios, no data from years prior to 2002 or from 2019 or 2020 are pooled. In both of these cases, analysts should use the variance structure on the released annual public use files. In the fifth scenario, only data from 2019 and 2020 are pooled, so analysts should use the variance structure on the released annual public use files. In the last scenario, data from year 2019-2020 are pooled with other years. The variance structure from HC-036 must be used in this circumstance as well. In no circumstance should the variance structure on the annual PUFs be combined with the variance structure on the HC-036 dataset.

The variables STRA9620 (stratum of the primary sampling unit) and PSU9620 (primary sampling unit) in this HC-036 dataset provide the appropriate sample design information needed by survey procedures in software packages that implement the with-replacement Taylor series linearization method to obtain estimates of complex sample variances.

The variables BRR1 – BRR128 in the HC-036BRR dataset (https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-036BRR) provide a comparable replicate sample design structure. These replicates can be incorporated in software package survey procedures that implement the balanced repeated replication (BRR) method to produce estimates of complex sample variances.

2.0 Data File Information

Released as an ASCII data file (with SAS®, STATA®, SPSS®, and R user statements) and in SAS Transport file, a SAS V9 file, a XLSX file, and a Stata file, the HC-036 file contains 442,093 records corresponding to the number of unique persons in MEPS from 1996-2020, with the exception of Panel 22 persons who appear in both the 2017 and 2018 HC files. The format for DUPERSID changed in 2018 requiring that HC-036 include these persons twice, one time with their 2017 DUPERSID and then again with their 2018 DUPERSID. All records contain the standard MEPS-HC person level ID variables (DUPERSID + PANEL), as well as the pooled variance estimation structure (STRA9620 and PSU9620).

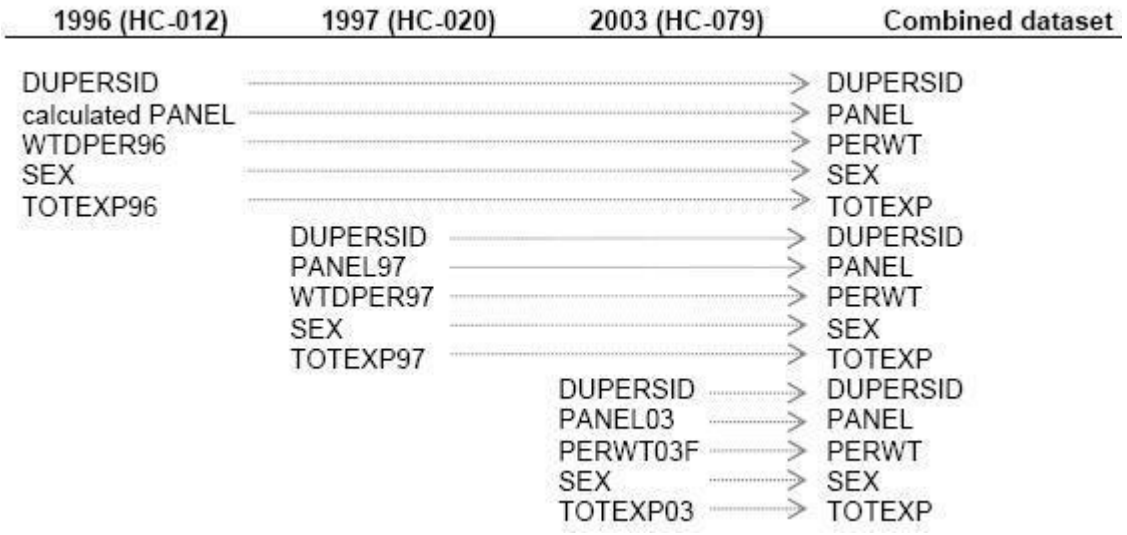
There is a record for each unique person appearing in any of the 1996-2020 MEPS HC full year person level public use files: HC-012, HC-020, HC-028, HC-038, HC-050, HC-060, HC-070, HC-079, HC-089, HC-097, HC-105, HC-113, HC-121, HC-129, HC-138, HC-141, HC-155, HC-163, HC-171, HC-181, HC-192, HC-201, HC-209, HC-216, and HC-224. These data sets have a combined total of 808,956 records; however, as each person may appear in one or two of these data sets, the number of records with unique DUPERSID in HC-036 (442,093) is fewer than the total number of records on the annual files.

3.0 Linking Instructions

The following steps should be taken to create a pooled analysis dataset.

1. Create a dataset for each year containing the person- and/or event-level records of all persons to be included in the analysis. Keep the unique person identifier (DUPERSID and PANEL), the person-level sampling weight, any classification variables (e.g., sex, race/ethnicity) and response variables (e.g., total expenditure amount, number of prescription drug purchases, etc.) to be used in the data analysis.

2. Reconcile the discrepancies in variable names. For all years, most variable names on the annual public use files contain a 2-digit year suffix. For instance, in the 1997 consolidated person-level file (HC-020) the panel variable is called PANEL97, the total annual expenditure amount variable is called TOTEXP97 and the sampling weight variable is called WTDPER97. But in the 2003 dataset (HC-079) these same variables are named PANEL03, TOTEXP03 and PERWT03F, respectively, and in the 1996 dataset (HC-012) the total expenditure and sampling weight variables are named TOTEXP96 and WTDPER96, respectively, and the panel variable is missing (users should assign a value of 1 for each record in HC-012). Starting in 2005, the panel variable is simply named PANEL (no year suffix). As illustrated below, the variable names must be made consistent before pooling the data.



3. Create a pooled analysis dataset by simply combining the individual-year datasets (e.g., the records from the 1996 and 1997 files). In other words, the number of records in the pooled file will equal the sum of the record counts for the individual annual files being pooled.

4. Attach the pooled variance structure to the pooled analysis dataset by merging the variables STRA9620 and PSU9620 from this HC-036 file to the pooled analysis dataset by DUPERSID and PANEL keeping all records in the pooled analysis dataset only. Depending on the software being used to manage the datasets, the pooled analysis dataset may need to be sorted by DUPERSID and PANEL prior to merging. This step will add two additional variables to the pooled file (STRA9620 and PSU9620) but have no impact on the number of records.

4.0 Adjustment of Analytic Weight Variable

It is generally recommended that analysts adjust the analytic weight variable by dividing it by the number of years being pooled. The sum of these adjusted weights represents the average annual population size for the pooled period (rather than the sum of the population sizes across multiple years that would result from unadjusted weights). Although this adjustment will have no effect on estimated means, proportions or regression coefficients because the weight variable is being divided by a constant (i.e. number of years), estimates of totals based on adjusted weights will reflect an "average annual" basis rather than the entire pooled period. On the other hand, if the objective is to produce an estimated total for the entire pooled period (e.g. total medical expenditures across multiple years rather than average per year), then the analytic weight variable should not be divided by the number of years in the pooled period.

5.0 Subpopulation Analysis Caveat

When pooling data over several years to increase sample sizes for small subdomains of the population (e.g., obtaining the total and mean expenditures for prescription drugs among children with asthma), users must be careful to maintain the integrity of the MEPS survey design. The MEPS design is accounted for by the full set of survey stratum and PSU values on both the annual files and this HC-036 pooled linkage variance estimate file.[1] When users create analytic subfiles that contain only respondents in the subdomain of interest (e.g., children with asthma), it is very unlikely that there will be all combinations of stratum and PSU that properly account for the MEPS survey design in a linearized estimate of the sampling variances. Therefore, the following approach is recommended for analyzing subpopulations in MEPS:

1) Construct a flag variable for all survey respondents that can be used to identify persons in the subdomain of interest,

2) Using a with-replacement design option for a Taylor Series procedure in a complex survey design statistical software package, read in records from all respondents (i.e., not just those in the subdomain of interest) and specify the analytic subdomain using the flag variable (see step 1 above).[2]

6.0 Further Information

For any question regarding the HC-036 file or pooling of data, please contact Sadeq Chowdhury by email at: sadeq.chowdhury@ahrq.hhs.gov or Fred Rohde by email at: frederick.rohde@ahrq.hhs.gov.

1 The MEPS design is also accounted for by the full set of replicates in the HC-036BRR data set (https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-036BRR).
2 The syntax for specifying survey designs and analytic subdomains varies across software packages (see section IB at https://meps.ahrq.gov/mepsweb/survey_comp/clustering_faqs.jsp for examples).