MEPS

Methodology Report #26

Variance Estimation from MEPS Event Files

Agency for Healthcare Research and Quality
U.S. Department of Health & Human Services

September 2011

ABSTRACT

A series of public use files (PUFs) are released from the Medical Expenditure Panel Survey (MEPS) Household Component every year. The most commonly used files are at the person level (with household and family identifiers) but some files are at lower levels such as medical conditions and medical events files. Eight event-level files are published from MEPS each year. For each person in the MEPS sample, these event files may include no record, a single record, or multiple records depending on the number of events the person had during the year. Therefore, an event file only includes records related to a subset of persons in the full year person file. Since the number of persons in an event file is not known before conducting the survey, any analysis of estimates from event files should be treated as a domain analysis which requires the entire sample to take all variability into account in estimating the variance of domain estimates. That is, the analysis should ideally include all persons with and without events in the file. However, in practice, since it is convenient to deal with the subset of cases with events only, users generally compute event-level estimates without merging persons with no event from the full person-level file. The impact of not doing a domain analysis is usually negligible if the subset (the domain) is large compared to the full file. This report looks into the issue of variance estimation from the MEPS event files and evaluates the impact of not doing a proper domain analysis on variance estimates.

Suggested Citation

Chowdhury S., Machlin S. *Variance Estimation from MEPS Event Files*, Methodology Report No. 26. September 2011. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr26/mr26.pdf

* * *

The estimates in this report are based on the most recent data available at the time the report was written. However, selected elements of MEPS data may be revised on the basis of additional analyses, which could result in slightly different estimates from those shown here. Please check the MEPS Web site for the most current file releases.

Center for Financing, Access, and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
www.meps.ahrq.gov

Background

The Medical Expenditure Panel Survey (MEPS)
Background

The Medical Expenditure Panel Survey (MEPS) is conducted to provide nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. MEPS is cosponsored by the Agency for Healthcare Research and Quality (AHRQ), formerly the Agency for Health Care Policy and Research, and the National Center for Health Statistics (NCHS).

MEPS comprises three component surveys: the Household Component (HC), the Medical Provider Component (MPC), and the Insurance Component (IC). The HC is the core survey, and it forms the basis for the MPC sample and part of the IC sample. Together these surveys yield comprehensive data that provide national estimates of the level and distribution of health care use and expenditures, support health services research, and can be used to assess health care policy implications.

MEPS is the third in a series of national probability surveys conducted by AHRQ on the financing and use of medical care in the United States. The National Medical Care Expenditure Survey (NMCES) was conducted in 1977, the National Medical Expenditure Survey (NMES) in 1987. Beginning in 1996, MEPS continues this series with design enhancements and efficiencies that provide a more current data resource to capture the changing dynamics of the health care delivery and insurance system.

The design efficiencies incorporated into MEPS are in accordance with the Department of Health and Human Services (DHHS) Survey Integration Plan of June 1995, which focused on consolidating DHHS surveys, achieving cost efficiencies, reducing respondent burden, and enhancing analytical capacities. To accommodate these goals, new MEPS design features include linkage with the National Health Interview Survey (NHIS), from which the sample for the MEPS-HC is drawn, and enhanced longitudinal data collection for core survey components. The MEPS-HC augments NHIS by selecting a sample of NHIS respondents, collecting additional data on their health care expenditures, and linking these data with additional information collected from the respondents' medical providers, employers, and insurance providers.

Household Component

The MEPS-HC, a nationally representative survey of the U.S. civilian noninstitutionalized population, collects medical expenditure data at both the person and household levels. The HC collects detailed data on demographic characteristics, health conditions, health status, use of medical care services, charges and payments, access to care, satisfaction with care, health insurance coverage, income, and employment.

The HC uses an overlapping panel design in which data are collected through a preliminary contact followed by a series of five rounds of interviews over a two and a half year period. Using computer-assisted personal interviewing (CAPI) technology, data on medical expenditures and use for two calendar years are collected from each household. This series of data collection rounds is launched each subsequent year on a new sample of households to provide overlapping panels of survey data and, when combined with other ongoing panels, will provide continuous and current estimates of health care expenditures.

The sampling frame for the MEPS-HC is drawn from respondents to NHIS, conducted by NCHS. NHIS provides a nationally representative sample of the U.S. civilian noninstitutionalized population, with oversampling of Hispanics and blacks.

Medical Provider Component

The MEPS-MPC supplements and validates information on medical care events reported in the MEPS-HC by contacting medical providers and pharmacies identified by household respondents. The MPC sample includes all hospitals, hospital physicians, home health agencies, and pharmacies reported in the HC. Also included in the MPC are all office-based physicians:

- Providing care for HC respondents receiving Medicaid.
- Associated with a 75 percent sample of households receiving care through an HMO (health maintenance organization) or managed care plan.
- Associated with a 25 percent sample of the remaining households. Data are collected on medical and financial characteristics of medical and pharmacy events reported by HC respondents, including:
- Diagnoses coded according to ICD-9 (9th Revision, International Classification of Diseases) and DSMIV (Fourth Edition, Diagnostic and Statistical Manual of Mental Disorders).
- Physician procedure codes classified by CPT-4 (Current Procedural Terminology, Version 4).
- Inpatient stay codes classified by DRG (diagnosis related group).
- Prescriptions coded by national drug code (NDC), medication names, strength, and quantity dispensed.
- Charges, payments, and the reasons for any difference between charges and payments.

The MPC is conducted through telephone interviews and mailed survey materials.

Insurance Component

The MEPS-IC collects data on health insurance plans obtained through private and public sector employers. Data obtained in the IC include the number and types of private insurance plans offered, benefits associated with these plans, premiums, contributions by employers and employees, and employer characteristics.

Establishments participating in the MEPS-IC are selected through two sampling frames:

- A Bureau of the Census list frame of private-sector business establishments.
- The Census of Governments from the Bureau of the Census.

Data from the two sampling frames are collected to provide annual national and State estimates of the supply of private health insurance available to American workers and to evaluate policy issues pertaining to health insurance. Since 2000, the Bureau of

Economic Analysis has used national estimates of employer contributions to group health insurance from the MEPS-IC in the computation of Gross Domestic Product (GDP).

The MEPS-IC is an annual panel survey. Data are collected from the selected organizations through a prescreening telephone interview, a mailed questionnaire, and a telephone follow-up for nonrespondents.

Survey Management

MEPS data are collected under the authority of the Public Health Service Act. They are edited and published in accordance with the confidentiality provisions of this act and the Privacy Act. 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 and microdata files. Summary reports are released as printed documents and electronic files. Microdata files are released on CD-ROM and/or as electronic files.

Printed documents and CD-ROMs are available through the AHRQ Publications Clearinghouse. Write or call: 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, 540 Gaither Road, Rockville, MD 20850; 301-427-1406, or email MEPSProjectDirector@ahrq.gov.

AHRQ Publications Clearinghouse
Attn: (publication number)
P.O. Box 8547 Silver Spring, MD 20907
800-358-9295
703-437-2078 (callers outside the United States only)
888-586-6340 (toll-free TDD service; hearing impaired only)

To order online, send an email to: ahrqpubs@ahrq.gov.

Be sure to specify the AHRQ number of the document or CD-ROM you are requesting. Selected electronic files are available through the Internet on the MEPS Web site: http://www.meps.ahrq.gov/

For more information, visit the MEPS Web site or email mepspd@ahrq.gov.

Table of Contents

Variance Estimation from MEPS Event Files

Sadeq Chowdhury, PhD and Steven Machlin, MS

Introduction

The Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.), and employers across the United States. MEPS provides estimates of specific health services use by the U.S. civilian noninstitutionalized population, the payments for these services, sources of payment, and the cost and scope of health insurance of U.S. workers. MEPS has three components: the Household Component (HC), Medical Provider Component (MPC), and the Insurance Component (IC). The Household Component collects data from individual households and their members in selected communities across the United States, drawn from a nationally representative subsample of households that participated in the prior year's National Health Interview Survey (conducted by the National Center for Health Statistics). The data collected from households are supplemented by data from their medical providers collected in the MPC. The Insurance Component is a separate survey of employers that provides data on employer-based health insurance.

The MEPS Household Component (which will be generally referred to as MEPS hereafter) collects detailed information for each person in the household on demographic characteristics, health conditions, health status, use of medical services, charges and source of payments, access to care, satisfaction with health care, health insurance coverage, income, and employment. The panel design of the survey, which features five rounds of interviewing covering two full calendar years, makes it possible to determine how changes in individuals' health status, income, employment, eligibility for public and private insurance coverage, use of services, and payment for care are related.

MEPS data are available on the MEPS Web site in data tables, downloadable data files (person, job, event, or condition level), and interactive data tools, as well as in publications using HC data. The main public use files (PUFs) released from MEPS are the Full Year (FY) Consolidated file and related medical conditions and medical event level files. The FY file includes records at the person level with family and dwelling unit (DU) identifiers and provides individual level information on health status, socio-demographic characteristics, employment, insurance, access to care, and various other related data items. The conditions file provides detailed information about each heath condition reported by the household respondent. It includes records at the person/condition level. Each event file consists of a specific type of event for persons in the corresponding full-year consolidated file. In the conditions file or an event file, some persons may have multiple conditions/events within the year and thus have multiple records and other individuals may have no medical conditions or events and have no records on the particular file. The conditions file or an event file includes identifiers to link each event to the individual in the person-level file who had the condition or the event.

In analyzing MEPS event (or conditions) files, estimates are often produced from the event file without merging all person records from the full file (i.e., only using the subset of persons with events). For point estimates, since the persons with no events will have no contribution to event related estimates, there is no need to keep those persons in the file. However, for variance estimation, since the persons with events (or conditions) are a subset of all persons, the subset total in the population is not known and the sample size is random, the theoretically correct approach is to include all persons (with or without an event) in the file and variance estimates are produced by treating the persons with and without events as separate domains. This approach to analysis is called domain analysis. Domain analysis takes the variability into account by using the entire sample in estimating the variance of subgroup estimates. For more information about domain analysis, see Lohr (1999), Cochran (1977), and Fuller et al. (1989). The 'domain' statement in SAS survey procedures, 'subgroup' statement in SUDAAN, and similar features in other variance computation software are designed to correctly estimate variances in such situations. If the estimates are produced using the records in the event file only, i.e., using the subset of persons with events from the FY file, the variances may not be estimated correctly. The extent of deviation of the estimated variance from the correct estimate for this technically improper analysis depends on the size of the subset compared to the full file. If the size of the subset is large and the number of variance strata with singleton primary sampling units (PSUs) is small or zero, the impact on the variance estimates will generally be negligible. This study examines this issue of variance estimation from the MEPS event files, with particular focus on the impact on variance estimates of analyzing the subset of cases with events only without doing a domain analysis that incorporates persons with no reported medical events.

MEPS Event Files

Eight public use event files from MEPS are published each year: prescription medicines, dental visits, other medical expenses, hospital inpatient stays, emergency room visits, outpatient department visits, office-based medical provider visits, and home care. Each of these files is an event-level file consisting of specific types of events for persons in the corresponding full-year consolidated file. Table 1 summarizes the sizes of 2008 MEPS event files and brief descriptions of different event files are provided below.

In a *prescription medicines event file*, each record represents a unique prescribed medicine event that is reported by the household respondent as being purchased or otherwise obtained for a household member. The file contains an identifier for each unique prescribed medicine and information on the detailed characteristics associated with the event; selected Multum Lexicon variables; conditions, if any, associated with the medicine; the date on which the person first used the medicine; total expenditures and sources of payments; and types of pharmacies that filled the household's prescriptions.

The *dental visits event file* contains variables pertaining to household reported dental visits. The file includes the date of the dental event, type of provider seen, if the visit was due to an accident, reason for the dental event, whether or not medicines were prescribed, expenditures, and sources of payment.

The *other medical supplies event file* contains information on the purchase of and expenditures for medical equipment, supplies, glasses and other medical items purchased, and sources of payment. Each record in this file contains information for the whole calendar year for all items except glasses for which each record contains information for a data collection round.

The *hospital inpatient stays event file* contains characteristics associated with the hospital inpatient stay event, such as the date of the hospital inpatient stay, reason for the stay, types of services received, condition(s) and procedure(s) associated with the hospital inpatient stay, whether or not medicines were prescribed, expenditures, and sources of payment.

The *emergency room visits event file* contains characteristics associated with the emergency room visit, such as the date of the visit, types of care and services received, types of medicine prescribed during the visit, condition codes, expenditures, and sources of payment.

The *outpatient visits event file* contains characteristics associated with the outpatient visit data, such as the date of the visit, type of provider seen, type of care received, type of services provided, expenditures, and sources of payment.

The *office-based medical provider visits event file* contains characteristics associated with the office-based visit, such as date of the visit, type of provider seen, time spent with the provider, types of treatment and services received, types of medicine prescribed, condition codes, expenditures, and sources of payment.

The *home health event file* can be used to make estimates of the utilization and expenditures associated with home health care. The file contains monthly information on expenditures for home health visits, types of providers, types of services received, lengths of visits, reasons for the visits, expenditures, and sources of payment. Each record in this file represents a month of care.

Moreover, the above files include various ID variables that can be used to link events to individuals in the FY person-level file or to other events or conditions in those files.

Domain Variance Estimation

Estimates from sample surveys like MEPS are often produced for different subgroups or subpopulations into which the population can be divided. For example, estimates may be required for groups with different types of health insurance, or persons with and without a health condition or events, or for a particular ethnic group. These subgroups are called domains or subdomains of study. The interest may concentrate on a particular domain in which the persons have certain characteristics or events, e.g., in the analysis of a particular type of event in MEPS. In that case, point estimates can be produced from the domain but the full sample is required for variance estimation. This is called domain analysis.

If the file is subset to the domain of interest only there will be no problem in producing the point estimates such as mean, percentage, or total. However, the variances or standard errors of these point estimates may be computed incorrectly from the subsetted file because the subset may not contain the full sample design information or the share of the domain to the full population to compute the variance correctly. This is not a problem if the sample is selected separately from each domain and the domain size in the population is known and the weighting adjustment is made independently within each domain. When the sample is not selected independently within each domain and the size is not fixed, the sample size becomes random in repeated draws. Also, if the population total of the domain is not known and not benchmarked at the domain level, the variance of the estimate of a total of a variable (say, total expense) not only depends on the variance of the mean expense per person, but also the variance of the estimate of the total number of persons in the domain. The full file with all domains is required to compute the variance of the total or the proportion that belong to the domain. The estimate of the mean in this case becomes the case of an estimate of a ratio because both the numerator and the denominator of the mean are estimates and the variance of the mean needs to be correctly estimated by treating it as a ratio. For means, this complication can be avoided by assuming that the sample size for the domain is fixed over repeated draws of the sample of the same overall size. The problem is more complex for computing the variance of an estimate of total.

If a simple random sample of size n is selected from a population of size N and the sample randomly includes n_c units from total N_c units in domain c. Then $w_i = \frac{N}{n}$ is the sampling weight (in the absence of nonresponse) for the *i*th unit $(i = 1,2,\cdots,n)$ with $\sum_i^n w_i = N$. If the variable of interest is y then the population mean, $\bar{Y}_c$, for domain c can be estimated as

$$\hat{\bar{Y}}_c = \frac{\sum_{i\in c}^n w_i y_i}{\sum_{i\in c}^n w_i},$$

and the population total, Y_c, for domain c can be estimated as

$$\hat{Y}_c = N_c\hat{\bar{Y}}_c = \sum_{i\in c}^n w_i y_i \quad \text{if } N_c \text{ is known}$$

$$= \hat{N}_c\hat{\bar{Y}}_c \qquad \text{if } N_c \text{ is unknown}$$

$$= \left(N * \frac{\sum_{i\in c}^n w_i}{\sum_i^n w_i}\right) * \frac{\sum_{i\in c}^n w_i y_i}{\sum_{i\in c}^n w_i} = \sum_{i\in c}^n w_i y_i$$

This shows for the point estimation of the mean and total for the domain, only the cases within the domain are required irrespective of whether the domain total N_c is known or unknown. Of course when N_c is unknown, it is implicitly estimated from the full sample. However, for variance estimation for domain estimates, since the sample size in the domain, n_c, is random, using only the cases within the domain is not sufficient to capture all components of the variance to compute the variance correctly. Moreover, since $\hat{N}_c$ is implicitly estimated from the sample for the estimate of total, the full sample is required to capture the variance of this component to accurately compute the variance of the estimate of total. In this case, the variance of the total is estimated from the full sample as follows:

$$\widehat{var}(\hat{Y}_c) = \frac{N^2 s^2}{n} \quad \text{ignoring finite population correction (fpc)}$$

where, $s^2 = \frac{1}{n-1}\sum_i^n (y_i^* - \bar{y}_i^*)^2$ with $y_i^* = \begin{cases} y_i, & if\ i \in c \\ 0, & otherwise \end{cases}$ and $\bar{y}^* = \frac{\sum_i^n y_i^*}{n}$. If the

variance is calculated from the cases in the domain only, it will not reflect the full variance of the estimate.

In addition to the theoretical reasons, there are practical reasons for keeping the full file and using the domain option for estimating the variance of a domain estimate. When a complex cluster sample design is used, the variance of a survey estimate is often computed based on variance strata and clusters (PSUs) using the Taylor series approximation. This approach needs at least two PSUs within each variance stratum to compute the variance by accounting for the variance contributions from all strata. If the domain of interest is small or clustered in certain areas so that some PSUs do not include any case from the domain, then some variance strata appear to have no PSU or only one PSU when the file is subset to the domain. The situation of one PSU within a stratum is known as the singleton PSU problem. In this case, it is not possible for variance computation software to correctly compute the variance from that stratum unless the full file is provided and a domain analysis is requested. In the absence of the full file, different software packages deal with a singleton PSU differently. SAS complex survey procedures exclude the singleton PSUs from variance calculation, SUDAAN imputes a value equal to the overall mean of all other PSUs for the missing PSU if the MISSUNIT option is used, and STATA offers different options including the approach SAS and SUDAAN use. Therefore, variance estimates from different software may not be identical. When there are more than two PSUs in a stratum and the domain has cases in at least two PSUs but not in all PSUs, none of the software can account for the missing PSUs when the file is subset to the domain. This can also lead to an underestimation of variances. Calculations of degrees of freedom (*df*), design effect, hypothesis testing, etc. are also affected by this. For example, to compute *df*, SUDAAN counts the number of PSUs and strata with at least one observation from the domain. In contrast, SAS 9.1 survey procedures compute *df* as the number of clusters (PSUs) in the non-empty strata minus the number of non-empty strata after excluding the singleton PSUs. When the *df* is not correctly computed it may affect the confidence interval or hypothesis testing and the resulting inference, particularly when the available *df* is small.

Theoretically, the variance is generally underestimated if the full file is not used. But the impact of a singleton PSU may be positive or negative depending on the situation and the software. However, overall the impact of using the subset and not the full file depends on the size and clustering of the domain compared to the full population. As the domain size gets larger, the impact becomes smaller and smaller both from theoretical grounds and because of the smaller number of singleton PSUs.

For further information about variance estimation for domain estimates, see Lohr (1999), Cochran (1977), Fuller et al. (1989), and user manuals for SAS survey procedures (SAS, 2004) and SUDAAN (Shah et al., 1997).

Comparison of Variance Estimates from Event Files

Table 1 presents a comparison of MEPS event files for 2008 in terms of the two factors that may affect the variance estimation from an event file—the file size and the number of variance strata with singleton PSUs. Three files with one or more singleton PSUs are Home Health, Hospital Inpatient Stays, and Outpatient Visits. These files also have the smallest number of persons. To investigate the impact on variance estimates when using the file subset to persons with events only, we compared the variances for selected estimates from these three files. Since these files have the smallest number of persons and have one or more singleton PSUs, any impact on variance from not doing a proper domain analysis should be more pronounced on the estimates from these files.

Table 1. Sizes of different MEPS Event files in 2008

Event file	Total number of event records	Corresponding number of persons and percentage of the full person file		Number of variance strata with single PSU
		Number	Percentage	
A. Prescribed Medicine	293,379	17,969	57.5	0
B. Dental Visits	26,253	11,639	37.2	0
C. Other Medical Expense	6,787	5,251	16.8	0
D. Hospital Inpatient Stays	2,821	2,113	6.8	4
E. Emergency Room Visits	6,115	4,165	13.3	0
F. Outpatient Visits	11,173	3,967	12.7	1
G. Office-Based Medical Provider Visits	136,460	21,208	67.8	0
H. Home Health	4,372	692	2.2	50

For the purpose of the analysis, each of these event files is merged with the 2008 FY person file by dwelling unit-person identifier (DUPERSID) and all records with necessary variables from both files are kept on the merged file. The merged file becomes an expanded event level file with one record for each person with no event (with missing values for event related variables) but one or more records for persons with events depending on the number of events. An indicator variable (say, event indicator) is created to indicate if the record came from the event file or not.

Estimates and standard errors (SEs) are then produced using SAS survey procedures in two different ways: 1) by subsetting only the records with events i.e., using a ‘by’ statement in SAS and 2) by performing a domain analysis using ‘event indicator’ as the domain. Using the ‘by’ statement only the persons with events are included in the analysis, which is equivalent to using the event file. In contrast, the expanded file with all person records with and without an event is used and estimates are produced when using the domain statement in SAS.

Tables 2 to 4 present comparative results under the two approaches for a selection of estimates of percentages, means, and totals from the three selected event files. As explained previously, the point estimates are the same under both approaches and the differences are only in standard errors (SEs). The differences in SEs are more pronounced for the estimates from the Home Health file which has 692 person records and 50 singleton PSUs, negligible for the estimates from the Hospital Inpatient Stays file which has 2,113 person records and only four singleton PSUs, and also negligible for the Outpatient Department visits file which has 3,967 person records and only one singleton PSU. For example, the SE of the estimate of mean expense per month for home health is 82.05 when produced from the event file and 85.38 when produced from the full file with the domain statement. For hospital inpatient stays, the SE for the estimate of mean expense per stay is 22.09 when produced from the event file and 22.15 when produced from the full file. For outpatient department visits, the SE for the estimate of mean expense per visit is 51.75 when produced from the event file and 51.66 when produced from the full file. A similar analysis was done using the Emergency Room Visits file, which has 4,165 person records and no singleton PSU, and not surprisingly found no difference in SEs of almost all estimates. It appears that the difference in SEs decreases or disappears as the number of persons increases and the number of singleton PSUs decreases in the event file.

When there is a difference, SEs are generally higher when estimated from the full file with domain analysis than when estimated from the event file. The differences are slightly more pronounced for SEs of totals than for SEs of means and percentages. There is a big difference in degrees of freedom (*df*) for estimating SEs from the event file and the full file. This is because in the event file some PSUs have no records but in the full file all PSUs have some records with or without events. However, since *df* is large in both cases, this difference in *df* will not have any impact on the inference here. If the *df* were small (say, less than 30) in one or both cases, the inference in terms of statistical testing or forming confidence intervals would be more precise from the full file.

As mentioned earlier, the treatment of singleton PSUs is different in SAS and SUDAAN. In SAS, singleton PSUs are excluded from the estimation of variances but in SUDAAN, when the MISSUNIT option is used, the overall mean of PSUs is used for the missing PSU to compute variances from a stratum with a singleton PSU. Since the number of stratum with singleton PSUs is small for the Inpatient and Outpatient files, there was no noticeable difference between the SE estimates from SAS in tables 3–4 and those from SUDAAN (not shown in any table).

Table 5 presents a comparison when the estimates are produced for subgroups within the Inpatient and Outpatient event files to see if there is any higher difference in SEs at that level. For this comparison, SEs are produced using three approaches: 1) subsetting the event file to the records in the subgroup of interest (i.e., using a ‘by’ statement in SAS), 2) using the event file and specifying subgroups as a domain, and 3) using the full file (with and without events) and specifying subgroups as a domain. The table shows that when a domain analysis is done either using the event file or the full file, the differences in SEs are small and negligible. However, if the analysis is done by subsetting the file to the subgroup of interest or by using a ‘by’ statement in SAS, the differences in SEs are substantial. For example, the SE of the estimate of mean expense for hospital inpatient stays for Hispanics is $1,295.6 when the analysis is done using the ‘by’ statement, $1,363.8 when the domain statement is used in the event file, and $1,360.6 when the domain statement is used in the full file. There are substantial differences in *df*. The *df* is substantially smaller when the estimates are produced by subsetting to the subgroup than when the estimates are produced using the domain statement either from the full event file or from the full file. The *df* available for variance estimation is large enough in either the full event file or the full file that the difference can be ignored. That means, if the domain statement is used for subgroup analysis, either the event file or the full file can be used without worrying about a substantial impact on the estimates of SEs or *df* for all event files except for the Home Health file. A ‘by’ statement or further subsetting of file to the subgroup of interest should never be used for analyzing any subgroup within an event file. In this comparison, the Home Health file is not included as the SEs are showing differences even at the overall level. Therefore, a domain analysis with the full Home Health file should always be used for estimation either at the overall or at the subgroup level.

Table 2. Comparison of variance estimates obtained from Home Health Event file and from merging of Event and Full Year files

Variable	Estimate[1]	SE of estimate		Degrees of freedom	
		Event file	Full file	Event file	Full file
Insurance status					
Private	29.76%	2.70	2.86	106	205
Public	69.49%	2.72	2.88		
Uninsured	0.75%	0.18	0.20		
Race/ethnicity					
Hispanic	9.93%	1.70	1.78	106	205
NH-black	16.83%	1.49	1.68		
Provider work for agency, hospital, nursing home?					
Yes	74.64%	2.42	2.52	106	205
Any care due to hospitalization?					
Yes	35.33%	2.50	2.59	106	205
Expense ($)					
Mean/month	$1,366	82.05	85.38	106	205
Total	$48.87B	4.72B	4.99B		
OOP expense ($)					
Mean/month	$153.7	37.17	37.62	106	205
Total	$5.49B	1.54B	1.55B		

[1] Since each record in Home Health file represents a month, the denominator for all percentage estimates is month of visits.

Table 3. Comparison of variance estimates obtained from Hospital Inpatient Stay Event file and from merging of Event and Full Year files

Variable	Estimate	SE of estimate		Degrees of freedom	
		Event file	Full file	Event file	Full file
Insurance status					
Private	54.58%	1.65	1.66	179	205
Public	40.27%	1.60	1.61		
Uninsured	5.15%	0.60	0.60		
Had surgery	39.20%	1.34	1.35	179	205
Race/ethnicity					
Hispanic	10.58%	1.03	1.03	179	205
NH-black	12.79%	1.01	1.02		
Expense					
Mean/visit	$11,349	427.11	424.59	179	205
Total	$329.9B	17.43B	17.73B		
OOP expense ($)					
Mean/visit	$312.6	22.09	22.15	179	205
Total	$9.09B	0.669B	0.675B		
Number of nights					
Mean/visit	5.22	0.21	0.21	179	205
Total	151.8M	8.46M	8.58M		

Table 4. Comparison of variance estimates obtained from Outpatient Department Event file and from merging of Event and Full Year files

Variable	Estimate	SE of estimate		Degrees of freedom	
		Event file	Full file	Event file	Full file
Insurance status					
Private	66.07	2.82	2.82	191	205
Public	30.48	2.90	2.90		
Uninsured	3.45	0.51	0.51		
Any surgery?					
Yes	12.24	0.78	0.78	191	205
Race/ethnicity					
Hispanic	6.08	0.77	0.77	191	205
NH-black	10.48	1.45	1.45		
Expense ($)					
Mean/visit	$792.67	51.75	51.66	191	205
Total	$97.76B	7.23B	7.25B		
OOP expense ($)					
Mean/visit	$66.60	5.28	5.28	191	205
Total	$8.21B	593M	595M		

Table 5. Comparison of variance estimates obtained for expense estimates from Event File and from merging of Event and Full Year files for subpopulations

		SE of estimate			Degrees of freedom		
	Estimate	Event file		Full file[1]	Event file		Full file[1]
		Subset	Domain[2]	Domain[2]	Subset	Domain[2]	Domain[2]
Expense: Hospital Inpatient Stay							
Mean for Hispanics	$12,081	1,295.6	1,363.8	1,360.6	57	179	205
Mean for blacks	$10,557	887.0	1,011.5	1,003.4	49		
Total for Hispanics	$37.2B	5.73B	6.14B	6.16B	57		
Total for blacks	$39.3B	3.18B	4.37B	4.46B	49		
Expense: Outpatient Department Visits							
Mean for Hispanics	$694	76.73	79.83	79.87	79	191	205
Mean for blacks	$734	63.19	72.15	72.20	74		
Total for Hispanics	$5.24B	627M	665M	666M	79		
Total for blacks	$9.06B	1,025M	1,310M	1,311M	74		

[1] Merging of the event file and full person file

[2] Domain analysis within the whole file

Conclusion

In estimating variances of estimates from MEPS event files, theoretically the event file should be merged with the FY person file and then a domain analysis should be used. This is required to account for the extra variance due to the fact that the number of persons in an event file (i.e., the sample size) is random and the corresponding population size is unknown. However, the impact of not doing a domain analysis on variance estimates is usually small when the subgroup is large and the number of singleton PSUs is small. To assess the impact on variances of producing estimates from MEPS event files without merging with the FY person file, an analysis is performed using the four smallest event files and the SEs of some estimates are compared.

The analysis shows that SEs are somewhat distorted (about 5 percent) for the Home Health file if the estimates are not produced from the full file with the domain option but this is not a notable problem for the other event files. Generally, the differences in variances between full and subsetted files are slightly higher for the estimates of totals than for means and proportions. There are differences in *df* available for estimating variances but the difference is ignorable since the available *df* under both approaches is sufficient. For estimating variances of estimates of subgroups within an event file, the differences in variances are negligible whether the estimates are produced from the full event file or the full file as long as the subgroup is treated as a domain. However, if the domain analysis is not done, the estimates of variance can be considerably biased and the available number of *df* can be less than sufficient.

In summary, for analysis of estimates from the Home Health file, the estimates of variances should always be computed by merging the event file with the full file with domain option. For all other event files, the analysis can be done using the event file only (i.e., without merging with the full file) without having any noticeable impact on the estimates of variances or *df*. For analyzing subgroups within the event file, a domain option should always be used and the file should never be further subset to subgroup of interest or the 'by' statement should never be used.

However, the above conclusion is based on the sizes of 2008 event files and this conclusion will be valid as long as the MEPS sample size (and hence the sizes of event files) remain stable from year to year. If there is a substantial decrease in the overall sample size, this conclusion may not be applicable and the caveats described above may need to be extended to other event files than just Home Health.

Finally, the reasons and the need for domain analysis discussed in this report are also applicable for analyzing any subset of a full person file. Generally, a domain analysis should be used for analyzing subgroup estimates using the full person file unless the impact of subsetting the file is assessed. This is particularly important when the subgroup size is not large and may be clustered geographically.

References

Botman S. L., Moore T. F., Moriarity C. L. Parsons V. L. *Design and Estimation for the National Health Interview Survey*, 1995–2004. National Center for Health Statistics. Vital Health Stat 2(130). 2000.

Cochran W. G. (1977). Sampling Techniques. New York, John Wiley & Sons, Inc.

Ezzati-Rice T. M., Rohde F., Greenblatt J. (2008). *Sample Design of the Medical Expenditure Panel Survey Household Component, 1998–2007*. Methodology Report No. 22. March 2008. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr22/mr22.pdf

Machlin S. R., Chowdhury S. R., Ezzati-Rice T., DiGaetano R., Goksel H., Wun L.-M., Yu W., Kashihara D. *Estimation Procedures for the Medical Expenditure Panel Survey Household Component.* Methodology Report #24. September 2010. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr24/mr24.pdf