Paper 1565-2014
Analyzing U.S. Healthcare Cost and Use with SAS®

Paul Gorrell, IMPAQ International LLC

ABSTRACT

A central component of discussions of healthcare reform in the U.S. are estimations of healthcare costs at the national level, as well as for subpopulation analyses for individuals with certain demographic properties or medical conditions. For example, a striking but persistent observation is that just 1% of the U.S. population accounts for more than 20% of total healthcare costs, and 5% account for almost 50% of total costs. In addition to descriptions of specific data sources underlying this type of observation, we demonstrate how to use SAS® to generate these estimates and to extend the analysis in various ways; that is, to investigate costs for specific subpopulations. The goal is to provide SAS programmers and healthcare analysts with sufficient data-source background and analytic resources to independently conduct analyses on a wide variety of topics in healthcare research. For selected examples, such as the estimates above, we concretely show how to download the data from federal web sites, replicate published estimates, and extend the analysis. An added plus is that most of the data sources we describe are available as free downloads.

INTRODUCTION

The featured article in the October 13, 2013 Washington Post Health & Science section was titled, "The 1% Solution" and described the cost to hospitals of, "the 1 percent whose ranks no one wants to join: the costly cohort battling multiple chronic illnesses who consumed 21 percent of the nearly $1.3 trillion Americans spent on healthcare in 2010." This is an enduring issue in discussions of healthcare reform, in particular discussions of healthcare cost and use. The January 2012 edition of the Atlantic Monthly contains an article with the following headline: "5% of Americans Made Up 50% of U.S. Health Care Spending." The article begins, "When it comes to America's spiraling healthcare costs, the county's problems begin with the 5%." But what is the basis for this, and similar, assertions about U.S. healthcare costs and use? The goal of this paper is to provide an introduction to the relevant data sources and SAS programming techniques needed to generate the types of estimates that are important for in- forming discussions of healthcare systems and reform.

Reading a bit further into the Atlantic Monthly article you see that the data source cited is the Agency for Healthcare Research and Quality (AHRQ). Each year AHRQ releases public-use files based on the Medical Expenditure Panel Survey (MEPS). In addition to these public-use files AHRQ publishes related reports and Statistical Briefs. The data and graphs in the Atlantic Monthly article were taken from a MEPS Statistical Brief, "The Concentration and Persistence in the Level of Health Expenditures over Time: Estimates for the U.S. Population 2008-2009." After introducing basic properties of the MEPS data file used to generate these estimates, we will show by use of example SAS code how to reproduce these estimates, as well as how to extend the analyses in various ways. Although we will focus on MEPS data, the goal is to illustrate general properties of healthcare data files and provide information which will allow analysts and SAS programmers to make productive use of the wealth of healthcare-related information released by the U.S. government each year.

GENERAL PROPERTIES OF HEALTHCARE DATA FILES

Although there is no substitute for thoroughly reading the documentation for each data source and data file (examples are given below to illustrate this), there are general properties that healthcare data files tend to share that can serve as a general guide. Before discussing these properties it's worth pointing out that documentation for publicly-released data files often include a documentation file which describes the data collection methodology and the various types of variables on the data files, as well as analytic guidelines. In addition there is a codebook file which lists formatted frequencies for each variable. Taking a close look at codebook frequencies is a prerequisite for a thorough understanding of the variable values on the data files. Most of the examples of variable values in the remainder of this section come from published MEPS codebooks.

Although this does not exhaustively represent the full array of information, it's useful to categorize the variables on healthcare data files as follows:

- ID (key)
- Demographic
- Health status and condition
- Utilization
- Expenditure
- Insurance coverage
- Population weight
- Variance estimation

Healthcare data files can be structured at different levels: person-level, family-level, event-level, condition-level, etc. A person-level file is one where each row uniquely contains information for a particular person, e.g. the person's age, gender, race/ethnicity, as well as medical conditions, insurance coverage, etc. An event-level file is one where each row uniquely contains information about a particular medical event, e.g. a visit to a doctor's office or a hospital stay. Although this type of file is actually at the person-event level, i.e. each row uniquely contains information for a particular condition for a specific person, it is most often referred to as an event-level file. Similarly a condition-level file is usually a person-condition level file, although one can imagine a condition-level file where each row contains summary, population-level, information for particular conditions (e.g. total costs for diabetes or congestive heart failure in a population during a particular time period).

ID variables, which often serve as key variables on data files, correspond to these various levels. For example, the Medical Expenditure Panel Survey (MEPS) and similar data files include a person-level ID variable that can be used as a merge variable when linking to other data files from the same data source. The MEPS event-level files contain an event ID (composed of the person-level ID variable plus a unique event identifier); the condition file has a condition ID (composed of the person-level ID variable plus a unique condition identifier).

Demographic variables, which are most often included on person-level files, include age, gender, socio-economic status (sometimes categorized as individual or household income as a percent of the Federal poverty level), geography, and race/ethnicity.

Health status variables may include reported or perceived health status, chronic condition indicators, as well as specific condition variables. Condition variables may be flag variables (e.g. diabetes: YES/NO) or ICD-9 or condition category codes at various levels (fully specified ICD-9 codes are often suppressed on public-use data files to preserve confidentiality).

Utilization variables, i.e. variables which indicate a person's use of the healthcare system, are usually specified by type of medical event, e.g. a visit to a doctor's office, a visit to an outpatient facility for an x-ray, or an inpatient hospital stay). Normally these are count variables for a specified time period, e.g. number of visits to a doctor's office in a particular year.

Expenditure variable values are normally actual costs for particular events, rather than provider charges. That is, they indicate the amount of money which was paid by a person or their insurance for a particular medical event. Expenditure variables may be further individuated by source of payment. For example, the cost for a doctor's visit may include both an out-of-pocket amount (a co-pay) as well as the amount paid by the insurance company.

Insurance coverage variables indicate what type of health insurance a particular person had at a particular time, or if they are uninsured. For example, a person with Medicare may also have additional private insurance. Some data sources such as MEPS provide monthly indicators which allow for tracking insurance status over time within the survey period.

Survey data are valuable for making population estimates. Population weight variables are used to generate population estimates at all available levels of analysis (person, family, event, condition). Because surveys often include oversampling, and weight variables account for non-response and other factors, it is important not to confuse sample frequencies or percents with population estimates. Proper use of the survey weight variables is essential for valid population estimates. Even within a particular survey you may need to pay particular attention to which weight variable is appropriate for your analysis. Surveys supplements or distinct components often have their own weight variable. We will see examples of the use of the weight variable for population estimates below.

National healthcare surveys are not simple, random, surveys, but (for various reasons) involve geographic clustering. To account for the correlations associated with this clustering, survey data are released with variance es-

timation variables to generate valid indicators of the reliability of the point estimates and statistically significant differences. Failure to account for the complex (non-random) nature of the survey design can lead to overestimation of the reliability of the point estimates. As with the weight variable, we will see examples of their use below.

Among many others, one important reason to read the data file documentation, including codebooks, is to gain a thorough understanding of the variable properties of each file. Intuition, or even past experience, is a dangerous guide when it comes to predicting the meaning of various variable values. For example, numeric variables based on survey data may have what I call pseudo-numeric values. Below I list a few common pseudo-numeric values used with survey data, along with their meaning (often indicated by the use of SAS formats).

VALUE	MEANING
-1	INAPPLICABLE
-7	REFUSED
-8	DON'T KNOW
-9	NOT ASCERTAINED

The negative values listed above indicate various types of survey non-response. A value of -1 (INAPPLICABLE) is often used to indicate a response that is inapplicable given a skip pattern in the survey (e.g. not asking a male respondent a question about pregnancy). Although these variable values are clearly indicated in the data file documentation, they can throw off any arithmetic computation with numeric variables. For example, MEPS medical event files contain numeric expenditure variables (costs for medical events) that also have "-1" values. If you look at the codebook for these data files you will see a list of variable values such as the following (where continuous positive values have been categorized by use of a format):

```
-1
0
$0.07   - $39.00
$39.01  - $74.53
$74.54  - $139.08
$139.09 - $24,000.00
```

The reason I call the "-1" value a pseudo-numeric value is because, although it IS a number, it does not have a numeric meaning. You clearly do not want this value entered into any computation, e.g. computing a mean, so you need to either delete these values or re-code them to zero (depending on your analytic needs).

A similar situation occurs with top coding. Top coding is the process where values on the high end of a range are grouped together to preserve confidentiality. Age variables are often top coded at 85 because comparatively few sampled individuals are older than 85. But this means that the numeric value "85" does not mean "85"; rather it means "85 or over". That is, it's a pseudo-numeric value.

A different type of example, but one which argues just as strongly for a thorough reading of documentation and related information such as codebooks, concerns changes to the meaning of variable values over time. For example the following illustrates changes to the correct interpretation of the values for the MEPS demographic variable RACEX after data year 2001.

VALUE	1996-2001	2002-2010
1	AMERICAN INDIAN	WHITE, NO OTHER RACE REPORTED
2	ALUEUT, ESKIMO	BLACK, NO OTHER RACE REPORTED
3	ASIAN, PACIFIC IS.	AMERICAN INDIAN / ALASKA NATIVE
4	BLACK	ASIAN
5	WHITE	NATIVE HAWAIIAN / PACIFIC IS.
6	---	MULTIPLE RACES REPORTED

Here the numeric values stay the same, but their meaning changes. If, as is not unusual, you were using MEPS data to track trends over time and had been using a standard program and formats for each new data year, the 2002 data based on this variable would have brought quite a surprise when looking at population estimates for the various groups listed above. That is, the program for 2002 data would not have generated any errors, but would

have indicated a significant change in the relative population size for, for example, the American Indian group. Aside from those changes, the only clue you might have had from the program output that something was off would have been the new numeric value "6". The lesson here is that it is not only important to read the documentation when first working with a particular data source, but also for each new data file you are using in your analysis.

In the next section we use MEPS data to look more closely at the process for generating estimates of U.S. healthcare costs and use. But the general points made below about generating population estimates apply to the Other Data Sources for the U.S. Population below, as well as numerous others not referenced there.

The Medical Expenditure Panel Survey (MEPS)

MEPS has a number of different components but in this paper we will focus on the Household Component (HC). The MEPS-HC (subsequently just referred to as MEPS) is a nationally representative survey of the U.S. civilian, non-institutionalized, population. It collects medical expenditure data as well as information on demographic characteristics, access to health care, health insurance coverage, and income and employment data. MEPS is co-sponsored by the Agency for Healthcare Research and Quality (AHRQ) and the National Center for Health Statistics (NCHS). Each year AHRQ releases a set of public-use files (PUFs). We will focus here on data files for calendar year 2009 since we want to show how to replicate estimates based on 2009 data. As is typical for publicly-release data, there is roughly a 2-year time lag between data year (e.g. 2009) and year of data release (e.g. 2011). This is due to a number of factors, e.g. the need to process end-of-year information in the next calendar year, as well as the analysis and processing required to create the population weights needed for analysis.

MEPS is an annual survey of roughly 35,000 people. What makes it valuable as a resource for analysis is that it is nationally representative. The person-level weight variable allows analysts to use the survey data for population and subpopulation estimates. We'll see the use of the weight variable in the SAS code examples below.

For the initial estimates reported here we will use the MEPS 2009 Full Year Consolidated Data File (H129). This is a public use file available for download from the MEPS web site (http://www.meps.ahrq.gov). See also the MEPS Factsheet "Computing Standard Errors for MEPS Estimates", also available from the MEPS web site. MEPS is not a simple, random, sample. Its design includes stratification, clustering, multiple stages of selection, as well as disproportionate sampling. MEPS public use files include variables for generating weighted national estimates and for use of the Taylor method for variance estimation. For 2009 these variables are: person-level weight (PERWT09F); stratum (VARSTR); and cluster/psu (VARPSU). We will not discuss variance estimation in any detail, but see Chen and Gorrell (2004), as well as the SAS documentation for the SAS survey procedures. In the next section we'll review the relevant properties of PROCs SURVEYMEANS and SURVEYFREQ since we will be using this procedure to generate population estimates with standard errors that take into account the MEPS complex design.

Among the set of annual MEPS PUFs are the following:

- Full-Year Consolidated Data File
- Medical Conditions File
- Medical Event Files
    - Office-based medical provider visits
    - Prescribed medicines
    - Outpatient visits
    - Emergency room visits
    - Hospital inpatient stays
    - Other medical expenses
    - Dental visits

If you go to the MEPS Data Files search page (accessed by clicking on the Data Files link in the left menu on the home page (http://meps.ahrq.gov). There is a lot of useful information available via these left-menu links, e.g. survey background, data overview, as well as Publications Search. This latter link will allow you to download the MEPS Statistical Briefs described below.

Appendix A shows the result of first clicking on the Data Files link in the left menu on the home page, and then searching for the 2009 Full Year Consolidated Data File (used in examples below). This page has links for the

documentation file, the codebook file, as well as the data file in either ASCII or SAS transport format. The txt file which contains the SAS Programming Statements includes the code you need to create a SAS data set from either the ASCII file or the SAS transport file you download, as well as label and format statements.

The full-year consolidated data file is a person-level file, i.e. as noted above, one row uniquely represents data for a single individual. It is a very wide file: the 2009 file (H129) has 1,881 variables. The full-year file, as well as the other MEPS PUFs, has a person ID variable which allows the various files to be linked. Because the full year file contains so many variables, it is usually a good idea to create an analytic file specific to your research goals by reading in only the variables you need.

The medical conditions file is a person-condition level file, i.e. each row uniquely represents information about a condition the person has reported. A person may have zero, one, multiple rows on this file, depending on the number of reported conditions for the year.

Each of the seven medical event files is at the person-event level where each row uniquely represents information associated with a particular medical event, e.g. an office visit or a hospital stay. Similar to the medical conditions file, a person will have zero, one, or multiple rows on an event file, depending on what events they have reported for the year.

The full-year file itself contains a wealth of information, although linking to the medical conditions or event files is often necessary to get detailed information about a person's condition or events, e.g. expenditure data for a particular event such as a hospital stay is found on the hospital inpatient stays file. But the full-year file is the best starting point for analyses using MEPS. We'll use this in the examples in the remaining sections of this paper.

The full-year file contains numerous types of variables:

- Demographic (e.g. age, sex, race/ethnicity)
- Utilization (e.g. counts of different medical event types)
- Expenditures (e.g. amount paid for medical events)
- Health insurance coverage (e.g. uninsured, covered by a private plan, Medicare, etc.)
- Sources of payment (e.g. which insurance type (private, Medicare) paid what amount)

This is only a partial list but serves to illustrate how much information is available for health care analysis. Of course you can combine these different types of information to look at healthcare cost and use for men compared with women or different age groups, or persons with different types of insurance coverage. Again, we will show examples below.

PROCs SURVEYMEANS AND SURVEYFREQ

SAS has a set of survey PROCs which are roughly comparable to their non-survey counterparts, e.g. PROC SURVEYMEANS is used with complex survey data for the same reason you would use PROC MEANS for non-survey data. Similarly for PROC SURVEYFREQ and PROC FREQ. The added functionality of the survey analysis procedures primarily concerns variance estimation. Because national healthcare surveys have a complex design, i.e. survey respondents are not randomly selected from among the U.S population, but rather drawn from geographic clusters or from oversampled subpopulations), the correlations among these groups must be taken into account when computing standard errors. Below I give a brief overview of the syntax of PROC SURVEYMEANS. For more detail see the SAS documentation. Chen and Gorrell (2008) has a more detailed introduction to the SAS survey analysis PROCs (based on SAS 9.1.3).

The following example illustrates the basic syntax and statements for PROC SURVEYMEANS (we'll come back to SURVEYFREQ below). What is unique to the survey procedures are the CLUSTER and STRATA statements. Depending on the survey and its documentation a CLUSTER is sometimes referred to as a primary sampling unit (psu). This is the case in MEPS and the CLUSTER variable is called VARPSU. The strata variable is VARSTR.

As noted above MEPS is a nationally representative sample. To generate population estimates based on the survey respondents you need to use the weight variable. For the 2009 MEPS files, this is PERWT09F (the final, person-level weight for the 2009 population). The population estimate is based on the Current Population Survey (CPS).

In the example below we are using PROC SURVEYMEANS to generate total and per-person average healthcare expenses for the 2009 civilian, non-institutionalized, population. The filename (H129) just is a sequence number attached to "H" (for household component). Note that all the examples in this paper can be replicated by downloading the MEPS data file from the MEPS Web site (see Appendix A) and using the listed SAS program code.

```
PROC SURVEYMEANS DATA= DMEPS.H129 SUM STD MEAN STDERR;
      WEIGHT PERWT09F;
      CLUSTER VARPSU;
      STRATA VARSTR;
      VAR TOTEXP09;
 RUN;
```

The options selected on the PROC SURVEYMEANS statement are SUM, STD (standard deviation of the sum), MEAN, and STDERR (standard error of the mean). The analysis variable listed on the VAR statement is the total expenditure variable for 2009. This variable represents each person's total expenditures for all reported medical events for the year. Running this code generates the following output (the ODS RTF destination was used).

Data Summary	
Number of Strata	165
Number of Clusters	370
Number of Observations	36855
Number of Observations Used	34920
Number of Obs with Nonpositive Weights	1935
Sum of Weights	306660588

Statistics				
Variable	Mean	Std Error of Mean	Sum	Std Dev
TOTEXP09	4107.003148	104.285565	1.259456E12	43229740433

In addition to the number of strata and clusters, the Data Summary table shows the total number of observations on the input data set (36,855), as well as the number of observations used (34,920). The reason for the difference between these two numbers is shown in the next row on the table: there are 1,935 observations with non-positive weights. The reason they are not used is because SURVEYMEANS deletes all observations with non-positive weights before running the analysis. This is because they are not relevant for generating population estimates since it's the weight variable that is used for this purpose. Of course this raises the question: Then why are these observations included on the file? The reason is that MEPS data files can also be used to generate family-level estimates (using a different weight variable), and a small number of individuals satisfy the criteria for being included in these estimates despite not satisfying the in-scope criteria for person-level estimates (see the MEPS documentation for a detailed discussion of this).

Note also the Sum of Weights in the last row of the table: 306,660,588. This is the CPS-estimated 2009 U.S. civilian, non-institutionalized, population.

The Statistics table shows the requested statistics. The SUM column shows that total expenditures for the entire MEPS population were $1.259 trillion in 2009. This estimate (with a bit of rounding) was published by AHRQ in its MEPS Statistical Brief #359 ("The Concentration of Health Care Expenditures and Related Expenses for Costly Medical Conditions, 2009"): "In 2009, health care expenses among the U.S. community population totaled $1.26 trillion."

The average expense per person, as shown in the Mean column is $4,104. Note that this average includes persons who had no reported healthcare expenses at all in 2009. Although it may not make a large difference when looking at a total population estimate, if you were looking more specifically at, for example, hospital inpatient stays where most people have zero expenditures, your estimate of the average cost of a hospital stay would be vastly underestimated due to all those zeroes. Even for the full population we can see the effect of restricting our analysis to those with an expense, i.e. not including those with zero expenditures for the year.

First a digression about subsetting. As Chen and Gorrell (2004) discussed, and detailed in the SAS survey procedure documentation, you need to avoid the normal inclination to subset to the observations you need for your analysis. This is because if you were to subset, for example, to those individuals with an inpatient stay in 2009, you would be deleting a large number of people and quite likely removing entire clusters from the analytic population. When you do this you prevent the procedure from knowing the full complex design properties and you risk generating invalid standard errors (often by underestimation, in my experience). So the rule of thumb is that it's great for efficiency reasons to delete variables (columns), but do not delete individuals, events, or conditions (rows). But then how do you just analyze the subpopulation you want?

First you need to create an indicator variable to identify the subpopulation of interest. For picking out individuals with a healthcare expense in 2009 we can use the following code:

```
IF TOTEXP09 > 0
      THEN SUBPOPX = 1;
ELSE SUBPOPX = 0;
```

This creates an indicator variable SUBPOPX with a value of 1 for individuals with an expense, and a value of zero for individuals without an expense. We can then use the following code:

```
PROC SURVEYMEANS DATA= DMEPS.H129 SUM STD MEAN STDERR;
      WEIGHT PERWT09F;
      CLUSTER VARPSU;
      STRATA VARSTR;
      VAR TOTEXP09;
      DOMAIN SUBPOPX;
RUN;
```

The DOMAIN statement allows for separate subpopulation analyses without deleting important survey design information. In addition to the Data Summary and Statistics tables above, the code above generates the following Domain Analysis table.

Domain Analysis: SUBPOPX						
SUBPOPX	Variable	N	Mean	Std Error of Mean	Sum	Std Dev
0	TOTEXP09	6772	0	0	0	0
1	TOTEXP09	28148	4854.960741	123.581319	1.259456E12	43229740433

Note that the sum for those where SUBPOPX=1 is the same as the total seen for the full population. This obviously follows from the fact that the only people not in this group are those with zero expenses. But now we see that the average is higher ($4,855 v. $4,107) since the denominator is smaller. Neither average is "wrong"; it just depends on what you want for your analysis.

Once you know the basic properties and variables in the input data, and understand the SURVEYMEANS procedure, you now have a large number of analyses that you can easily run. For example, if you wanted to know how average expenses for men and women compared, you could use the SEX variable with the DOMAIN statement, as below:

```
PROC SURVEYMEANS DATA= DMEPS.H129 SUM STD MEAN STDERR;
      WEIGHT PERWT09F;
      CLUSTER VARPSU;
      STRATA VARSTR;
      VAR TOTEXP09;
      DOMAIN SEX;
      FORMAT SEX SEXF.;
RUN;
```

The variable SEX on the data file is numeric, with 1=MALE and 2-FEMALE. The format used just makes this clear. Similar to what we saw with SUBPOPX, use of the SEX variable generates the following Domain Analysis table:

Domain Analysis: SEX						
SEX	Variable	N	Mean	Std Error of Mean	Sum	Std Dev
1 MALE	TOTEXP09	16634	3559.281882	120.206508	535983283788	20705962169
2 FEMALE	TOTEXP09	18286	4635.473201	171.517056	723472715012	32163478979

Here we see that the average is higher for women compared to men ($4,635 v. $3,559). This difference is also reflected in the subpopulation totals (Sum column): $723 billion for women and $536 billion for men.

Again, once you have a core set of basic information about the data, and experience with the relevant SAS procedures, it becomes comparatively easy to generate a wide variety of estimates. Using similar code you can look at average expenses for person under 65 com[pared to those 65 and over, as in the following code:

```
PROC SURVEYMEANS DATA= DMEPS.H129 SUM STD MEAN STDERR;
      WEIGHT PERWT09F;
      CLUSTER VARPSU;
      STRATA VARSTR;
      VAR TOTEXP09;
      DOMAIN AGE09X;
      FORMAT AGE09X AGEF.;
RUN;
```

The 2009 age variable (AGE09X) is a continuous numeric variable. Here the format groups values into 0-64 and 65 and over (it also distinguishes a small number with -1 [INAPPLICABLE] values). This code generates the following table:

Domain Analysis: AGE09X						
AGE09X	Variable	N	Mean	Std Error of Mean	Sum	Std Dev
-1 INAPP	TOTEXP09	247	14659	2011.587778	38724680169	6322734922
0-64	TOTEXP09	30919	3231.729999	105.229252	857147322416	36500513184
65+	TOTEXP09	3754	9373.056681	327.433935	363583996215	18503766971

Not surprisingly we see that the average per-person expenditure is higher for those 65 and over v. those under 65 ($9,373 v. $3,231).

Although not a factor here it's important to note that, as mentioned above, the age variables on MEPS files (as well as numerous other public-use files) have top-coded age values. On MEPS files age is top-coded at 85. As noted, this means that a variable value of "85" means '85 or over', not '85'. Top coding is done to preserve confidentiality for small subpopulations. It is always important to thoroughly read data documentation so that you understand the data's strengths and limitations before running analyses. For example, you cannot use the age variable to compute average age (especially for those over 65) because the top-coded value of 85 is not really a numeric value.

Now we're in a position to return to the estimates from the beginning of the paper: in 2009 1% of the U.S. population accounted for more than 20% of total healthcare costs, and 5% account for almost 50% of total costs. We're going to use SURVEYMEANS again but we need to set things up first by defining the top 1% and 5% of the population. These will be defined by their total healthcare expenses in 2009 (TOTEXP09). So first we need to sort the dataset DESCENDING by expenses:

```
PROC SORT DATA= DMEPS.H129;
      BY DESCENDING TOTEXP09;
RUN;
```

We're interested in population estimates so we need the weighted top 1% and 5%, so now that the dataset is sorted so that those with the largest annual expenses are at the top of the file, we can create a new weight variable CUMWT which is a cumulative population weight variable. This will allow us to capture percent of the population based on expenditures. With a properly sorted dataset it's easy to create this variable:

```
DATA DMEPS.H129;
      SET DMEPS.H129;
      CUMWT+PERWT09F;
RUN;
```

Now we can use the CUMWT variable to create two new variables, one for the top 1% (TOTEXPT1) and one for the top 5% (TOTEXPT5). Here the value of SUMTOT is the sum of the weights we saw in the last row of the Data Summary table (306,660,558). Therefore the statement IF CUMWT <= (SUMTOT*.01) picks out the weighted top 1% of the population, and TOTEXPT1=TOTEXP09 creates a variable TOTEXPT1 that is the total healthcare expenditure (TOTEXP09) for everyone in this group. Similarly for TOTEXPT5=TOTEXP09 for the top 5%.

```
IF CUMWT <= (SUMTOT*.01)
      THEN TOTEXPT1=TOTEXP09;
ELSE TOTEXPT1 = 0;
IF CUMWT <= (SUMTOT*.05)
      THEN TOTEXPT5=TOTEXP09;
ELSE TOTEXPT5 = 0;
```

OK, now we have the variables we need to run PROC SURVEYMEANS and generate our percent-of-total-expenditures estimates for the top 1% and top 5% of the population. The RATIO statement is very handy for this. The denominator is the total expenditure variable (TOTEXP09) and the denominator is either the expenditure variable for the top 1% (TOTEXPT1), as in the example below, or the expenditure variable for the top 5%.

```
PROC SURVEYMEANS DATA= DMEPS.H129 SUM;
      WEIGHT PERWT09F;
      CLUSTER VARPSU;
      STRATA VARSTR;
      VAR TOTEXPT1 TOTEXP;
      RATIO TOTEXPT1 / TOTEXP09;
RUN;
```

Recall that the Atlantic Monthly article stated that, "In 2009, the top 1% of patients accounted for 21.8% of expenditures". In the Ratio Analysis table below we see this as 21.7961%

Ratio Analysis			
Numerator	Denominator	Ratio	Std Err
TOTEXPT1	TOTEXP09	0.217961	0.018896

If we were to run the same code substituting the top 5% variable (TOTEXPT5), we would generate the following:

Ratio Analysis			
Numerator	Denominator	Ratio	Std Err
TOTEXPT5	TOTEXP09	0.495050	0.012913

This is the basis for the headline that 5% of Americans made up 50% of U.S. health care spending (49.505%).

Once you have identified the top 1% and 5% of the population with respect to expenditures, a natural set of follow-up questions concern healthcare utilization, i.e. what medical events are driving these costs? We can take an initial look at this by using three MEPS utilization variables:

- OBTOTV09
- IPDIS09
- OPTOTV09

These utilization variables are continuous variables that are counts of particular medical events in 2009 (the '09' suffix). OBTOTV09 is a count of office-based provider visits (OBV), e.g. a typical visit to the doctor. IPDIS09 is a count of inpatient hospital discharges (IPD), i.e. number of inpatient hospital stays. OPTOTV09 is a count of outpatient visits, i.e. a visit to an outpatient clinic or other provider for a lab test, x-ray, etc.

We can use PROC SURVEYFREQ to see how the top 1% compare to the rest of the population with respect to these particular medical events. To do this we first need an indicator variable that picks out the top 1%. We can use the expenditure variable we created above to do this. Recall that TOTEXPT1 only has a positive value for persons in the top 1%. Given this the code below will create our indicator variable:

```
IF TOTEXPT1 > 0
      THEN TOP_01 = 1;
ELSE TOP_01 = 0;
```

If you have used PROC FREQ then you already have a jumpstart on using PROC SURVEYFREQ. The syntax is similar, with the obvious difference that you also have the CLUSTER and STRATA variables we saw used with SURVEYMEANS above. The TABLES statement is similar to PROC FREQ.

The code below outputs frequency and percent information for two subpopulations (those in the top 1% and those who are not) for office visits in 2009. The continuous OBTOTV09 variable has been formatted into 3 categories: those with zero visits, those with 1-5 visits, and those with 6 or more visits in 2009. The table on the next page was generated by the following code:

```
PROC SURVEYFRQ DATA= DMEPS.H129;
      WEIGHT PERWT09F;
      CLUSTER VARPSU;
      STRATA VARSTR;
      TABLES TOP_01*OBTOTV09 / ROW;
      FORMAT OBTOTV09 OBVISITS. ;
RUN;
```

There is a lot of useful information in this table, and I will focus on the percent columns below, bur we'll first walk through the rows and columns.

The rows are divided into three sections, one for each value of the TOP_01 indicator variable (Note that the table would have a different orientation if the TABLES statement had been OBTOTV09*TOP_01), and one for the overall total. Each section has its own total so you can look at the analysis for each subpopulation.

One advantage that SURVEYFREQ has over PROC FREQ is that the output tables contain both the unweighted and weighted frequencies. These are shown in the first two data columns. The (unweighted) Frequency column is useful for looking at sample size for each subpopulation cell. The overall unweighted frequency is 34,920. This is the Number of Observations Used we saw in the Data Summary table above. The overall Weighted Frequency Total (306,660,588) is the Sum of Weights in that Data Summary table, i.e. the CPS-estimated U.S. community population in 2009. The next column shows the standard deviation for the weighted frequency. Although beyond the scope of this paper, standard deviations and standard errors are crucial information for determining the reliability of the frequency and percent point estimates.

Note that, although not explicitly labeled as such, the Percent columns are weighted percents.

For our look at the office-based visit utilization pattern of the top 1% as compared with the other 99%, the Percent column is not particularly useful since the cell values in this column are percents of the total (the astute reader will note that the top 1% is shown here to really be the top 0.994%).

More useful is the Row Percent column. This is not generated by default by PROC SURVEYFREQ, but must be requested by using the ROW option on the TABLES statement, as shown in the code above. Here we can see the percent for each formatted value of OBTOTV09 within each subpopulation group (recall that TOP_01 = 1 for the top 1% and 0 for the remaining 99%).

The Row Percents reveal interesting differences between the two groups. For example, over 70% of the top 1% had 6 or more office visits in 2009, compared with 24% of the remaining 99%, and only 5% had no office visits, compared with 27% of the remaining 99%.

Table of TOP_01 by OBTOTV09

TOP_01	OBTOTV09	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
0	0	11084	82601786	2282016	26.9359	0.4434	27.2063	0.4486
	1-5	16422	146963643	3488011	47.9239	0.4034	48.4050	0.4049
	6 OR MORE	7104	74047097	1963118	24.1463	0.3996	24.3887	0.4024
	Total	34610	303612525	6634463	99.0060	0.0759	100.000	
1	0	19	161073	45536	0.0525	0.0148	5.2844	1.4675
	1-5	74	722511	121372	0.2356	0.0388	23.7039	3.2562
	6 OR MORE	217	2164479	189136	0.7058	0.0603	71.0116	3.2112
	Total	310	3048063	242584	0.9940	0.0759	100.000	
Total	0	11103	82762859	2283566	26.9884	0.4434		
	1-5	16496	147686153	3507000	48.1595	0.3987		
	6 OR MORE	7321	76211576	1998976	24.8521	0.4018		
	Total	34920	306660588	6698793	100.000			

To look at utilization patterns for hospital inpatient stays, we need only substitute IPDIS09 for OBTOTV09 in the TABLES statement in the SURVEYFREQ code above. For the table shown on the top of the next page, we used a different format to look at those with 0, 1, or more than 1 hospital stay in 2009.

Looking again at the Row Percent column, we see a striking contrast for inpatient hospital stays: the vast majority of the population (i.e. the 99% subpopulation we've defined) had no hospital stay in 2009 (94%). But over 80% of the top 1% had a hospital stay in 2009. Further (extrapolating a bit beyond the percents shown in the table), the weighted frequencies show that, among those in the 99% with a visit, fewer than 20% had more than one hospital visit; for the 1%, about 60% of those with a hospital visit had 2 or more.

Table of TOP_01 by IPDIS09								
TOP_01	IPDIS09	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
0	0	32450	284304916	6289627	92.7100	0.1868	93.6407	0.1777
	1	1803	16000875	581056	5.2178	0.1559	5.2702	0.1573
	2 OR MORE	357	3306734	229110	1.0783	0.0747	1.0891	0.0754
	Total	34610	303612525	6634463	99.0060	0.0759	100.000	
1	0	53	571206	105152	0.1863	0.0338	18.7400	3.2320
	1	105	999600	139001	0.3260	0.0444	32.7946	3.3289
	2 OR MORE	152	1477257	152683	0.4817	0.0494	48.4654	3.5375
	Total	310	3048063	242584	0.9940	0.0759	100.000	
Total	0	32503	284876121	6309075	92.8962	0.1914		
	1	1908	17000475	611520	5.5437	0.1613		
	2 OR MORE	509	4783991	286944	1.5600	0.0930		
	Total	34920	306660588	6698793	100.000			

We see a similar pattern when we look at outpatient visits, as shown in the table below.

Table of TOP_01 by OPTOTV09								
TOP_01	OPTOTV09	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
0	0	30481	261778667	5809543	85.3643	0.3786	86.2213	0.3676
	1	2399	23971846	942029	7.8171	0.2337	7.8955	0.2372
	2 OR MORE	1730	17862013	811368	5.8247	0.2443	5.8832	0.2468
	Total	34610	303612525	6634463	99.0060	0.0759	100.000	
1	0	150	1334140	141720	0.4351	0.0446	43.7701	3.3496
	1	44	409564	75268	0.1336	0.0246	13.4369	2.2577
	2 OR MORE	116	1304359	148601	0.4253	0.0477	42.7930	3.1648
	Total	310	3048063	242584	0.9940	0.0759	100.000	
Total	0	30631	263112807	5842964	85.7994	0.3722		
	1	2443	24381409	953530	7.9506	0.2373		
	2 OR MORE	1846	19166371	837432	6.2500	0.2499		
	Total	34920	306660588	6698793	100.000			

Here we see that over half (56%) of those in the top 1% had an outpatient visit, but only about 14% of those in the remaining 99% did.

With these basic tools in place, and given the range of variables available on the MEPS PUFs (as well as compa-

rable data files available from the data sources listed in the next section), there is an extensive range of analyses available with a few simple SAS programs. For example, where we have looked at subpopulation analyses of men compared with women, and those with high expenditures compared with those with lower expenditures, we could also look at race/ethnicity by using variables such as RACEX (shown above), or persons with particular conditions, e.g. by using the ICD-9 code values of the condition variables or specific disease indicators.

In addition, as noted above, MEPS has a set of source-of-payment variables which allow researchers to investigate what types of expenditures (e.g. for which event type) are associated with which types of insurance, or self-pay. We can illustrate this with the following example. The variable RXMCR06 is the expenditure variable for prescribed medicines (RX) paid for by Medicare (MCR) in 2006 (06). 2006 is an interesting year with respect to Medicare payments because it was the first year of implementation of Medicare Prescription Drug Coverage (Medicare Part D).

As we have above, we can use PROC SURVEYMEANS to investigate healthcare expenditures. Here, instead of the total expenditure variable, we'll use the 2006 variable that is specific to (i) prescribed medicines, and (ii) Medicare as a source of payment (H105 is the 2006 MEPS consolidated full year data file).

```
PROC SURVEYMEANS DATA= DMEPS.H105 SUM STD MEAN STDERR;
      WEIGHT PERWT06F;
      CLUSTER VARPSU;
      STRATA VARSTR;
      VAR RXMCR06;
RUN;
```

The Data Summary table below shows MEPS information for the 2006 population. Note that the U.S. civilian population estimate is 299,267,035 for 2006, compared to 306,660,588 for 2009.

Data Summary	
Number of Strata	203
Number of Clusters	451
Number of Observations	34145
Number of Observations Used	32577
Number of Obs with Nonpositive Weights	1568
Sum of Weights	299267035

Statistics				
Variable	Mean	Std Error of Mean	Sum	Std Dev
RXMCR06	148.156750	6.847958	44338431281	2136641644

The Statistics Table for 2006 shows that the average expenditure per prescribed medicine is $148, and the sum of all prescribed medicine expenditures is $44.339 billion. This estimate is reported in the AHRQ MEPS Statistical Brief #240 ("Prescription Drug Estimates for Medicare Beneficiaries, 2005 and 2006"), "Medicare expenditures for outpatient prescription drugs were more than seven times as high in 2006 than the prior year, rising from $5.9 to $44.3 billion." I hope it is clear how you would verify the 2005 estimate of $5.9 billion.

Of course you may be interested in seeing how Medicare expenditures have changed since 2006. You could use the 2009 file and replace the "06" variables with the "09" variables in the SURVEYMEANS code. Doing so would generate the Statistics table below.

Statistics				
Variable	Mean	Std Error of Mean	Sum	Std Dev
RXMCR09	202.005744	9.354779	61947200054	2962591109

Now we can see that total Medicare expenditures for prescription drugs for the non-institutionalized population increased from $44.3 billion in 2006 to $61.9 billion in 2009. In part this could be explained as a function of the increase in population from 2006 to 2009 (299 million to 306 million), but we can see from the increase in average per-person expenditures, from $148 in 2006 to $202 in 2009, that this is only part of the story.

Of course you now have the tools to investigate other sources of payment, and look at particular subpopulations. Suppose you were interested in the trend for self-pay (out-of-pocket) prescription drug expenditures from 2006 to 2009. Instead of using the Medicare source of payment variable (RXMCRyy) you would use the self-pay variable (RXSLF06, RXSLF09). You may also be specifically interested in the 65-and-over population since you may be investigating Medicare Part D's effect on self-pay. Using what you learned above you could use an age format to look at self-pay prescribed medicine costs for persons 65 and over in 2006 and in 2009.

The code below is for 2006, but it should be clear what minimal changes are needed to also run this with 2009 data.

```
PROC SURVEYMEANS DATA= DMEPS.H105 SUM STD MEAN STDERR;
      WEIGHT PERWT06F;
      CLUSTER VARPSU;
      STRATA VARSTR;
      VAR RXSLF06;
      DOMAIN AGE06X;
      FORMAT AGE06X AGEF. ;
  RUN;
```

As shown in the tables below, in 2006 the average, annual, per-person, self-pay expenditures for prescription medicines was $693. In 2009 this had decreased to $516. Total self-pay expenditures for prescription medicines decreased from $25.2 billion to $20.0 billion.

Domain Analysis: AGE06X					
AGE06X	Variable	Mean	Std Error of Mean	Sum	Std Dev
<65	RXSLF06	200.350343	8.536072	52656917059	2640359788
65+	RXSLF06	692.753709	20.039624	25245914631	1020322973

Domain Analysis: AGE09X					
AGE09X	Variable	Mean	Std Error of Mean	Sum	Std Dev
<65	RXSLF09	134.432887	4.535483	36010571570	1407615694
65+	RXSLF09	515.950145	20.705200	20013878266	1026960511

Of course for a thorough analysis of what is going on you'd also want to look at the under-65 population, as well as changes in other sources of payment, e.g. private insurance. You could use the RATIO statement to look at particular sources of payment as a percentage of the total.

But the goal of this paper has not been to answer all your questions concerning estimates of healthcare costs and use, but to give you the information needed to answer these questions on your own. The next section lists additional data sources, including crucial data file documentation and SAS program examples.

OTHER DATA SOURCES FOR THE U.S. POPULATION

All the publicly-released data files referenced below are available as downloads at no cost. It is important that you read and respect all data-use guidelines and agreements. In addition, be sure to read all relevant data file documentation before conducting any data processing or analysis. Data set variables have specific properties which must be understood in order to generate valid analyses.

The data sources listed below are only a small sample of healthcare data files released each year.

Behavioral Risk Factor Surveillance System (BRFSS)

- State-based system of health surveys that collects information on health risk behaviors, preventive health practices, and health care access primarily related to chronic disease and injury
- The world's largest telephone health interview survey
- Data years available online (1984 – 2011)
- General information: http://www.cdc.gov/brfss/index.htm
- Data file documentation and downloads: http://www.cdc.gov/brfss/technical_infodata/surveydata.htm
- SAS program information: Select a data year on the survey data page above and scroll down to the SAS Resources section

Medicare Claims Public Use Files

- The Centers for Medicare and Medicaid Services (CMS) releases basic standalone (i.e. non-linkable), de-identified, Medicare claims-based data files
- Provides claims-based information on healthcare cost and utilization for the Medicare population
- Data based on a 5% sample of Medicare beneficiaries
- General information: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/index.html
- SAS program information: For each data file there is a SAS data users guide with basic SAS code

National Health and Nutrition Examination Survey (NHANES)

- National survey which combines interviews with physical examination and lab tests
- The interview includes demographic, socioeconomic, dietary, and health-related questions
- The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by medical personnel
- General information: http://www.cdc.gov/nchs/nhanes.htm
- Data file documentation and downloads: http://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm
- Data files released every two years (1999-2000 to 2011-2012)
- Earlier releases available (NHANES I, II, and III)
- SAS program information: NHANES releases numerous data files in SAS transport format. For each set of files (e.g. Demographics, Dietary) SAS code examples for merging data files are provided.

National Health Interview Survey (NHIS)

- National household survey of non-institutionalized population. Main objective is to monitor health of the U.S. population
- Questionnaire consists of a core set of questions (which vary little year to year), as well as supplements on specific topics (e.g. child mental health, cancer control)
- Provides national estimates of health conditions, utilization, access to care, and health behaviors
- General information: http://www.cdc.gov/nchs/nhis.htm
- Data file documentation and downloads (1997 to present): http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm
- SAS program information: Select a data year on the link above, then the Data Release link, and then the

Sample SAS Statements link

SUMMARY

Each year the U.S. government releases a wealth of healthcare data that allow researchers to investigate issues related to cost and use, among numerous other topics. Many of these data files are made available online as free downloads in SAS transport format and with example program code to create SAS datasets. SAS is an important tool in this research and those with SAS programming knowledge are in an excellent position to conduct important research into current topics in health services research.

REFERENCES

Agency for Healthcare Research and Quality [AHRQ] (2009) MEPS HC-129 2009 Full Year Consolidated Data File documentation. Available for download at: http://meps.ahrq.gov/mepsweb/data_stats/download_data/pufs/h129/h129doc.pdf.

Cohen, S. and Yu, W. The Concentration and Persistence in the Level of Health Expenditures over Time: Estimates for the U.S. Population, 2008–2009. Statistical Brief #354. January 2012. Agency for Healthcare Research and Quality, Rockville, MD. Available for download at: http://www.meps.ahrq.gov/mepsweb/data_files/publications/st354/stat354.pdf.

Cohen, S. Statistical Brief #359. The Concentration of Health Care Expenditures and Related Expenses for Costly Medical Conditions, 2009. February 2012. Agency for Healthcare Research and Quality, Rockville, MD. Available for download at: http://www.meps.ahrq.gov/mepsweb/data_files/publications/st359/stat359.pdf.

Chen, X. and Gorrell, P. (2004) "Variance Estimation With Complex Surveys: Some SAS-SUDAAN Comparisons". NESUG 2004 Proceedings. Available for download at: http://www.nesug.org/Proceedings/nesug04/an/an02.pdf.

Chen, X. and Gorrell, P. (2008) "An Introduction to the SAS Survey Analysis PROCs". NESUG 2008 Proceedings. Available for download at: http://www.nesug.org/proceedings/nesug08/sa/sa06.pdf.

Stagnitti, M. N. Prescription Drug Estimates for Medicare Beneficiaries, 2005 and 2006. Statistical Brief #240. March 2009. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.gov/mepsweb/data_files/publications/st240/stat240.pdf.

Weissman, J. "5% of Americans Made Up 50% of U.S. Health care Spending". Atlantic Monthly, January 2012. Available online at: http://www.theatlantic.com/business/archive/2012/01/5-of-americans-made-up-50-of-us-health-care-spending/251402/.

ACKNOWLEDGMENTS

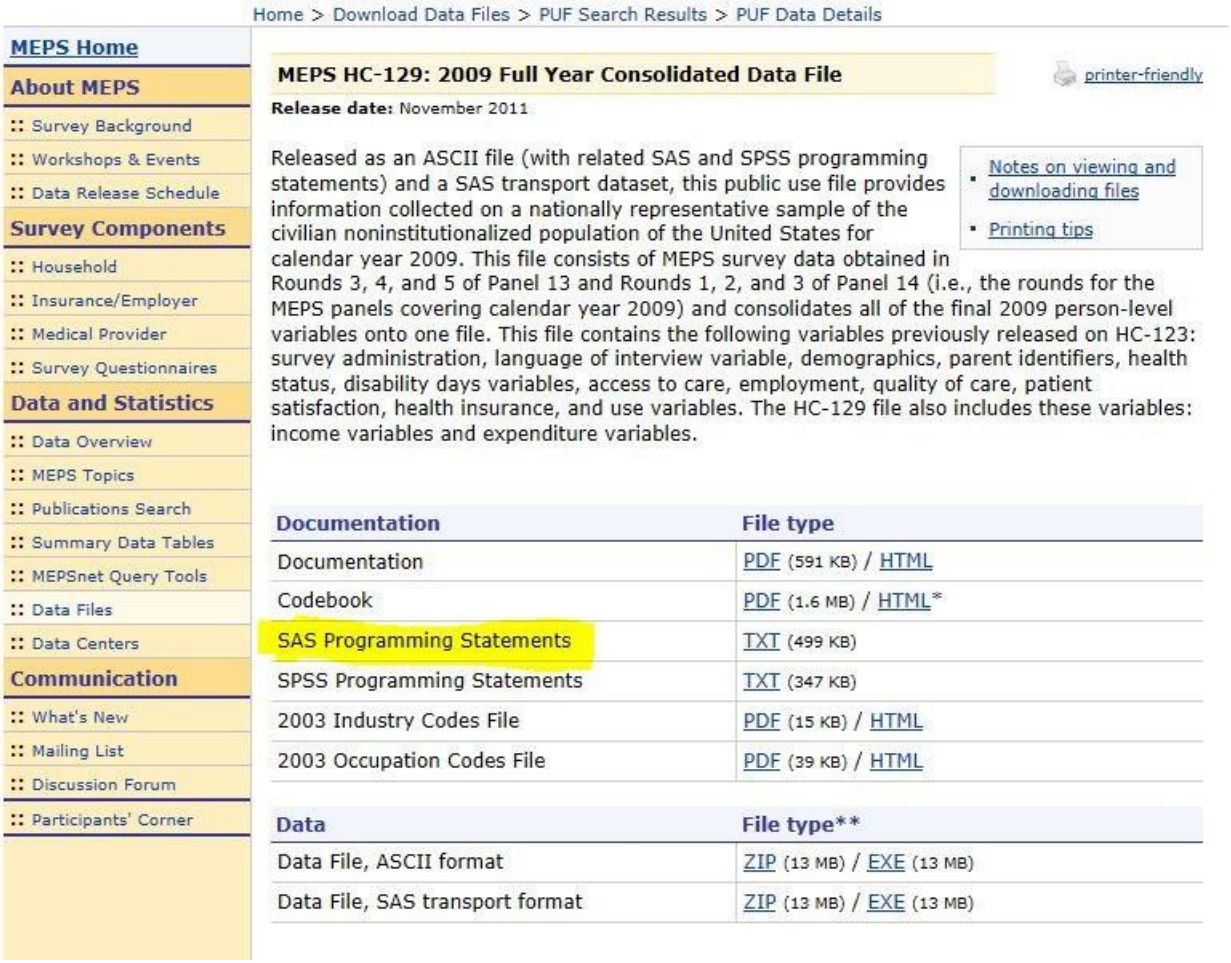
Thanks to Zhengyi Fang for his insights concerning the generation of the weighted percentages for the top 1% and 5% of the population with respect to healthcare expenditures.

CONTACT INFORMATION

I would appreciate any comments or feedback on this paper. Contact me at:

Paul Gorrell
IMPAQ International LLC
1101 Vermont Avenue, NW
Washington, DC 20005
pgorrell@impaqint.com

Appendix A
MEPS Web Site Screen Shot (2009 Full Year Consolidated Data File Page)
(http://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-129)



Home > Download Data Files > PUF Search Results > PUF Data Details

MEPS Home
About MEPS
:: Survey Background
:: Workshops & Events
:: Data Release Schedule
Survey Components
:: Household
:: Insurance/Employer
:: Medical Provider
:: Survey Questionnaires
Data and Statistics
:: Data Overview
:: MEPS Topics
:: Publications Search
:: Summary Data Tables
:: MEPSnet Query Tools
:: Data Files
:: Data Centers
Communication
:: What's New
:: Mailing List
:: Discussion Forum
:: Participants' Corner

MEPS HC-129: 2009 Full Year Consolidated Data File

printer-friendly

Release date: November 2011

- Notes on viewing and downloading files
- Printing tips

Released as an ASCII file (with related SAS and SPSS programming statements) and a SAS transport dataset, this public use file provides information collected on a nationally representative sample of the civilian noninstitutionalized population of the United States for calendar year 2009. This file consists of MEPS survey data obtained in Rounds 3, 4, and 5 of Panel 13 and Rounds 1, 2, and 3 of Panel 14 (i.e., the rounds for the MEPS panels covering calendar year 2009) and consolidates all of the final 2009 person-level variables onto one file. This file contains the following variables previously released on HC-123: survey administration, language of interview variable, demographics, parent identifiers, health status, disability days variables, access to care, employment, quality of care, patient satisfaction, health insurance, and use variables. The HC-129 file also includes these variables: income variables and expenditure variables.

Documentation	File type
Documentation	PDF (591 KB) / HTML
Codebook	PDF (1.6 MB) / HTML*
SAS Programming Statements	TXT (499 KB)
SPSS Programming Statements	TXT (347 KB)
2003 Industry Codes File	PDF (15 KB) / HTML
2003 Occupation Codes File	PDF (39 KB) / HTML

Data	File type**
Data File, ASCII format	ZIP (13 MB) / EXE (13 MB)
Data File, SAS transport format	ZIP (13 MB) / EXE (13 MB)