An Examination of State Estimates Using Multiple Years of Data from the Medical Expenditure Panel Survey, Household Component

John Sommers

Agency for Healthcare Research and Quality Working Paper No. 06004

May 2006

Suggested citation: Sommers J. An Examination of State Estimates Using Multiple Years of Data from the Medical Expenditure Panel Survey, Household Component. Agency for Healthcare Research and Quality Working Paper No. 06004, May 2006, http://gold.ahrq.gov.

AHRQ Working Papers provide preliminary analysis of substantive, technical, and methodological issues. The papers have not undergone external peer review. They are distributed to share valuable experience and research. Comments are welcome and should be directed to the authors. The views expressed are those of the authors and no official endorsement by the Agency for Healthcare Research and Quality or the Department of Health and Human Services is intended or should be inferred.

An Examination of State Estimates Using Multiple Years of Data from the Medical Expenditure Panel Survey, Household Component
John Sommers
May 2006

ABSTRACT

For smaller subsets of population, such as, states, the sample size from a single year sample from the Medical Expenditure Panel Survey (MEPS), Household Component (HC), is not enough to provide quality estimates for many variables. Because of the importance of these estimates non standard methods need to be developed to obtain estimates for these small subsets of the population. Two possible methods that can be considered are pooling of multiple years of data or use of small subpopulation estimation techniques.

This paper first examines the decrease in errors by using two or three consecutive years of data from the MEPS-HC rather than a single year. Due to sample overlap between years in the MEPS – HC, data from year to year are correlated and additional sample has less effect on total error than it would if the years had independent samples. The loss of sample efficiency depends on the correlations of the variables estimated.

In this paper a variety of estimates are produced and improvements in standard errors using multiple years are calculated and compared to the errors obtained using one year of data. The improvement in errors over single year estimates are compared to the theoretical decrease in errors that would be obtained if the data across years were uncorrelated.

The paper then examines the effect of using one type of small subpopulation estimation technique on the same variables. This technique is applied to earlier estimates constructed from one, two and three years of MEPS-HC data. Finally, the effects of using both multiple years and the special estimation technique are examined and reported.

John Sommers
Mathematical Statistician
Center for Financing, Access, and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
E-mail: John.Sommers@ahrq.hhs.gov

An Examination of State Estimates Using Multiple Years of Data from the Medical Expenditure Panel Survey, Household Component

Background

The demand for state estimates of health expenditures, health status, percentages of persons with certain health conditions, and uninsurance rates is growing. In a recent report (Sommers, 2005), the possibility of making state level estimates with data from the Medical Expenditure Panel Survey (MEPS), Household Component (HC) was examined. This work showed that for larger states many quality estimates can be made either directly or using some simple small area estimation procedures. However, the paper also showed that as the sample size decreased, either due to the size of the state or due the uncommon nature of the event, the size of the error for both the direct and the small area estimates increased to a size that was problematic. One can assume other estimates for smaller sub populations within a state would have worse results.

A common method to improve estimates is to pool several years of data. The Census Bureau produces poverty and income estimates using data from 2 or 3 years of data. (Census Website). This same technique can be applied to the types of estimates examined in the earlier work with the MEPS HC data.

This paper presents results using estimates made with HC data from the years 2001 through 2003. Combined year estimates are made for the identical variables as the previous report by Sommers, 2005, and estimates are also made for subpopulations of persons age 18 or older in the survey that were either diabetic, hypertensive, asthmatic, arthritic or obese. Although these groups overlap and the classification of persons into these groups is based upon self reported information and the results are still useful since the study is intended to explore the effects of using multiple years of data on estimated errors for small populations.

If samples of independent data are pooled it is a relatively simple matter to estimate the effects of using a larger sample. However, this is not the case if one pools data from the MEPS HC across years. Samples from different years of the MEPS HC are not independent. The HC has overlap in both persons and PSU's (Cohen, 2000). Thus, the values are correlated across years.

With correlation, if we had the same variance for each year of the HC and we were simply to average the same estimate for two years of the survey, then the variance for this estimate would be:

$$\text{var} = \left(.5*\left(\sigma^2\left(1+\rho\right)\right)\right)$$

where ρ is the correlation between the estimates for the two years. One can see as the correlation increases towards + 1 there are diminishing returns in this approach. If the correlation were 1, then the average of the two years would be no better than that for a single year. If the correlation were 0 then the result would be the equivalent of doubling the sample. It is unlikely that the correlation is negative since, for example, with an expenditure estimate; this would require persons who were the highest spenders one year to be lower than average the next. Research has shown that persons with high expenditures in one year tend to have high expenditures the following years. (Monheit, ref)

Similar formulas that depend upon the pariwise correlations between two years can be developed for averages across three years of data. In general, addition of each extra year with the same sample size, errors and equal correlations with the previously added years will result in a smaller reduction in error relative to the error for a single year. For instance, if there were no correlation between years, the standard error using two years of data would be .707 times the error of the one year sample. Using three years of data would yield an error of .577 times the error of the one year sample and four years would result in an error of .5 times the error of a single year sample. As one can see the reduction in error relative to the original year one of sample is less for each additional year of sample.

Although the MEPS HC does not follow this simplified model since samples and correlations for the MEPS HC are not equal for all years, this model gives an indication of how correlation can effect the reduction in errors of estimates using multiple years of data relative to the error of similar estimates made with a single year of data. Reduction of errors is decreased by positive correlation between years. Further, there is a diminishing effect on the reduction in error with the addition of each additional year of data. How, much extra reduction in error one gets using three years of the MEPS - HC rather than one or two is part of the focus of this work.

Estimates Analyzed

For this analysis, estimates were produced using three years of HC data, 2001, 2002 and 2003. Estimates were made for each single year, for 2001 and 2002 combined, 2002 and 2003 combined and for the combination of all three years. Since one of the key reasons for analysis is to determine the possibility of producing state estimates, estimates were made for each of the twenty largest states by population for each of five conditions for persons age 18 and over. The five conditions were obesity, diabetes, hypertension, asthma and arthritis. A person was classified as obese using their height and weight measurements from the survey. A person was classified as having one of the other four conditions based upon questions on the survey that asked if they had ever been diagnosed with the condition. For each state by condition by year’s combination, three different direct estimates were produced: the percent of persons with the condition, the percent of those persons with the conditions who had a health expenditure during the year and the conditional mean expenditure for those who had an expenditure. For each of these estimates, a corresponding small area estimate was made using a compositing technique similar to that used in previous work. (Sommers, 2005). These small area estimates have a relative mean squared error smaller than that of the standard direct estimate. These were made to determine the average effect of this process when applied to direct estimates for different numbers of years of data. Other types of estimates by type of expenditure were also made, but will not be discussed because the results relative to the total gain from using the same number of years of data or the small area technique are similar to the results that are reported in this paper.

Terminology

Before the analysis of the results is presented, it would be of value to establish some useful terminology. An important factor being considered is the ratio of errors for the same type of estimate, made with multiple years versus a single year of data. In the previous section, it was discussed how under ideal conditions with the same sample size and no correlations between years, how this value would be .707 for two years of data , .577 for three years of data and .500 for 4 years of data. This ratio of errors will be referred to in this paper as the multi year error ratio for the type of estimate. For a specific number of years, for example, for a two year estimate it will be called the two year error ratio. An average of this value over a set of estimates, say a set of states, will be referred to as the average error ratio. The values of .707, .577 and .500 will be referred to as the ideal error ratio.

Another term used will be state size group. There are two groups used: group 1 is the ten largest states by population, and group 2 is the second ten largest states by population.

Direct Estimates

Table 1 shows average rse's, relative standard errors, for the percent of persons with each of the five conditions for the single year 2003, for the combination of 2002 and 2003 and the combination of 2001 through 2003. The table also shows the average two and three year error ratios for cells defined by state size group and condition. The error used was the rse for each estimate. (Note, since the rse is the error over the mean and the means are about equal for the one, two and three year estimate, when error ratios are created with rse's they are very similar to the error ratios for standard errors. Rse's were used because of convenience.)

(Also, note that the average error ratios given in the tables are the average ratios of the rse's for one and either the two or three year estimate for the same item. Although many times this average of ratios is similar to the ratio of the average of all rse's for one year estimates and the average of the rse's for the set of two or three year estimates, they are not the same. Sometimes with sample sizes used in this paper they can be noticeably different.)

Although, the average rse's differ considerably by condition and state size group, the effect on the relative standard errors of using two and three years of data on the different groups of states and conditions is very similar for all the groups of estimates. The average two year error ratio for all estimates is slightly more than 84 percent. The average 3 year error ratio for all estimates is about 70 percent.

One can see the effect of the correlations across the years. As was discussed earlier, the estimated value of the two year and three year error ratios, if there were no correlations across years would be approximately 71 and 58, percent respectively. Since the average error ratios are significantly larger than these values, this is an indication of significant correlations of the percent that have any of these conditions across the years.

TABLE 1
Comparisons of Relative Standard Errors for Direct State Estimates of the Percent of the Population with Selected Conditions: Multiple Years versus a Single Year of Data: Adults Age 18 Years and Older

		COMBINATION OF YEARS				
CONDITION	STATE SIZE GROUP	2003	2002-2003		2001-2002-2003	
		Ave Rse	Ave Rse	Ave Two Year Error Ratio	Ave Rse	Ave Three Year Error Ratio
Arthritis	1	.076	.067	.867	.053	.717
	2	.098	.079	.818	.078	.720
Asthma	1	.121	.097	.831	.093	.704
	2	.161	.131	.817	.130	.719
Diabetes	1	.139	.121	.861	.101	.677
	2	.187	.153	.832	.145	.704
Hypertension	1	.061	.049	.824	.046	.706
	2	.086	.072	.870	.065	.700
Obesity	1	.079	.062	.852	.049	.667
	2	.105	.088	.860	.076	.708

Table 2 shows results for percents of persons who have any expense or a dental expense for 3 of the conditions for the two state size groups and a single year, two years and three years of data. These results are representative of the estimates of percents with expenditures.

For any of the conditions whether less prevalent, such as, diabetes, or more visible, such as, arthritis or obesity, the relative errors for the estimate of percent of persons who have any expense are very low. This is because the percents are very close to 100, almost all these individuals have expenditures every year, and a binomial variable with that probability of occurrence has a very small relative error. The percent, which have a dental expenditure, is of more interest. These estimates are closer to 50% and have more representative relative standard errors. These errors using one year of data, are very high for diabetes, because the sample size of persons with diabetes is small. The values for one year are better for persons with arthritis or who are obese. These are the more frequent in occurrence of the 5 conditions.

On the average for all estimates over all conditions, the two year error ratio is about 77%. The variation from 77 percent by condition or state size groups is not large. The comparable average three year error ratio is 65 percent.

These values are lower than those obtained for estimates of the percentages of the population with a particular condition. This indicates that the correlation of whether a person in the overlapping sample has a condition is higher than the chance that they will have an expenditure across the two years. However, there is still significant correlation across the years of data for whether a person with a particular condition has an expenditure for different types of medical care. This can be seen by comparing the average two and three year error ratios obtained for these estimates of the percentages that have expenditures, 77 and 65 percent with the comparable ideal error ratios of 70.7 and 57.7, percent respectively.

TABLE 2
Comparisons of Relative Standard Errors for Direct State Estimates of the Percent of the Population with a Dental or Medical Expenditure among Persons with Selected Conditions: Multiple Years versus a Single Year of Data: Adults Age 18 Years and Older

		COMBINATION OF YEARS				
CONDITION AND EXPENDITURE GROUP	**STATE SIZE GROUP**	**2003**	**2002-2003**		**2001-2002-2003**	
		Ave Rse	**Ave Rse**	**Ave Two Year Error Ratio**	**Ave Rse**	**Ave Three Year Error Ratio**
Arthritis						
Dental expenditure	1	.089	.072	.817	.058	.660

Dental expenditure	2	.133	.107	.821	.100	.697
Any expenditure	1	.013	.008	.745	.008	.626
Any expenditure	2	.012	.010	.772	.010	.667
Diabetes						
Dental expenditure	1	.199	.158	.752	.126	.620
Dental expenditure	2	.312	.208	.742	.185	.613
Any expenditure	1	.014	.009	.824	.008	.730
Any expenditure	2	.005	.004	.808	.004	.805
Obesity						
Dental expenditure	1	.100	.078	.778	.060	.595
Dental expenditure	2	.128	.115	.839	.101	.716
Any expenditure	1	.030	.019	.663	.016	.557
Any expenditure	2	030	.023	.750	.019	.615

Table 3 gives results for estimates of conditional mean expenditures for persons with representative conditions who had a dental or any expenditure. As is the general case with estimates for expenditures, the relative standard errors are much higher than those for estimates percent of persons who had an expenditure. This type of estimate is one of the main reasons that one would use multiple years of data to make an estimate. For diabetes, even with 3 years of data, the relative standard errors for conditional mean expenditures are very marginal in quality

The average value of the multiple year error ratios over all conditions and states for the average conditional mean expenditure are very similar to those for estimates the percentage of persons with an expenditure. They do vary by condition, expenditure type and state size group.

TABLE 3
Comparisons of Relative Standard Errors for Direct State Estimates of the Conditional Mean Expenditures for Dental and All Medical Expenditures for Persons with Selected Conditions: Multiple Years versus a Single Year of Data: Adults Age 18 Years and Older

		COMBINATION OF YEARS				
CONDITION AND EXPENDITURE GROUP	**STATE SIZE GROUP**	**2003**	**2002-2003**		**2001-2002-2003**	
		Ave Rse	**Ave Rse**	**Ave Two Year Error Ratio**	**Ave Rse**	**Ave Three Year Error Ratio**
Arthritis						
Dental expenditure	1	.160	.143	.767	.116	.655
Dental expenditure	2	.193	.177	.840	.144	.620
Any expenditure	1	.117	.097	.797	.087	.665
Any expenditure	2	.151	.124	.798	.111	.695
Diabetes						
Dental expenditure	1	.339	.259	.760	.259	.734
Dental expenditure	2	.331	.258	.816	.232	.733
Any expenditure	1	.156	.141	.777	.099	.652
Any expenditure	2	.283	.193	.768	.095	.622
Obesity						
Dental expenditure	1	.215	.144	.739	.116	.634
Dental expenditure	2	.252	.192	.797	.164	.665
Any expenditure	1	.168	.132	.832	.102	.655
Any expenditure	2	.165	.121	.754	.118	.651

Composite Estimators

As one can see, especially in Table 3, multiple years of data do not guarantee that the relative standard errors for the direct estimates will be less than 20 percent. For instance, average relative standard errors for estimates of conditional mean dental expenditures for persons with diabetes, average over 20 percent. For more common conditions, these same types of estimates have relative standard errors that average over 10 percent. Many other multi year estimates that we have examined also have this problem.

While one could add even more years of data, averaging over larger numbers of years could be problematic because of changing conditions. For instance, a state's ranking in health care costs could change over the years due to specific policies implemented by the state. Averaging over a large number of years might hide this change. Also, as we have seen, relative to the error obtained with one year of data, each additional year of data has a diminishing effect on the error.

To avoid using more years of data, a composite type of small area estimator similar to that used in earlier work, Sommers, 2005, was applied to the conditions examined. Using the methodology described in that work, composite estimates and mean squared errors were created for one, two and three years of data by shrinking the direct estimates towards the regional estimates. Because these estimates are biased relative to the state values, errors include estimates of variance and bias. Errors also include estimates of addition to variances caused by estimation of a weighting parameter required to build the composite estimate from the direct regional and state estimates.

Tables 4, 5 and 6 show the effects of this process on the estimates presented in Tables 1, 2 and 3. Errors used in these tables are the relative standard error for the unbiased direct estimates and the relative mean squared errors for the composite estimates. The latter estimates are made up of both variance and bias components. Note any improvement from this process is in addition to any improvement resulting from combining multiple years of data to make an estimate. In order to obtain estimates of overall average error ratios for estimates which combine the compositing methodology with addition of additional years of data, one must multiply the error ratios obtained from using multiple years of data shown in Tables 1, 2 and 3 times comparable error ratio values in Tables 4, 5 and 6 which show the effects of compositing. For instance, for states in group 1, from Table 1 one can see that use of two years of data for estimates of the percent of persons with arthritis yields an error that is .867 times the average for a single year. On table 4 for the same two year estimate the RSE is .843 times that of the direct estimates. Thus, the improvement using both two years of data and the composite methodology gives an estimate with a standard error that is about .731 times the error for a direct estimate made with a single year of data.

Although the numbers on the tables vary considerably, when the data are analyzed using an analysis of variance model, only one factor showed a significant effect on the prediction of the improvement due to using the composite estimation methodology. This factor was the state groups. The use of this methodology reduced errors more for the smaller states in state group 2 compared to the improvements for the larger states in state groups 1. Neither condition type, item estimated (percent with expenditure versus conditional mean); type of expenditure nor did number of years of data have a significant effect on the average error ratio of the composite estimate versus the direct estimate. On average this ratio was about .66. The average for the set of larger states was about .70, while the average for the smaller states was about .62.

TABLE 4
Comparisons of Relative Errors for Composite versus Direct State Estimates of the Percent of the Population with Selected Conditions Made with One, Two or Three Years of Data: Adults Age 18 Years and Older

		COMBINATION OF YEARS		
CONDITION	**STATE SIZE GROUP**	**2003**	**2002-2003**	**2001-2002-2003**
		Ave Error Ratio of Composite versus Direct Estimates	**Ave Error Ratio of Composite versus Direct Estimates**	**Ave Error Ratio of Composite versus Direct Estimates**
Arthritis	1	.831	.843	.848
	2	.628	.610	.540
Asthma	1	.655	.624	.654
	2	.544	.453	.428
Diabetes	1	.594	.601	.678
	2	.608	.689	.657
Hypertension	1	.705	.795	.777
	2	.658	.655	.586
Obesity	1	.759	.682	.814
	2	.490	.584	.560

TABLE 5
Comparisons of Relative Standard Errors for Composite versus Direct State Estimates of the Percent of the Population with a Dental or Medical Expenditure for Persons with Selected Conditions Made with One, Two or Three Years of Data: Adults Age 18 Years and Older

		COMBINATION OF YEARS		
CONDITION AND EXPENDITURE GROUP	**STATE SIZE GROUP**	**2003**	**2002-2003**	**2001-2002-2003**
		Ave Error Ratio of Composite versus Direct Estimates	**Ave Error Ratio of Composite versus Direct Estimates**	**Ave Error Ratio of Composite versus Direct Estimates**
Arthritis				
Dental expenditure	1	.760	.552	.835
Dental expenditure	2	.761	.746	.686
Any expenditure	1	.684	.726	.748
Any expenditure	2	.578	.684	.709
Diabetes				
Dental expenditure	1	.600	.601	.643
Dental expenditure	2	.671	.585	.526
Any expenditure	1	.641	.660	.653
Any expenditure	2	.612	.655	.528
Obesity				
Dental expenditure	1	.776	.783	.833
Dental expenditure	2	.644	.767	.675
Any expenditure	1	.618	.637	.664
Any expenditure	2	.841	.737	.695

TABLE 6
Comparisons of Relative Standard Errors for Composite versus Direct State Estimates of the Conditional Mean Expenditures for Dental and All Medical Expenditures for Persons with Selected Conditions Made with One Two or Three Years of Data: Adults Age 18 Years and Older

		COMBINATION OF YEARS		
CONDITION AND EXPENDITURE GROUP	**STATE SIZE GROUP**	**2003**	**2002-2003**	**2001-2002-2003**
		Ave Error Ratio of Composite versus Direct Estimates	**Ave Error Ratio of Composite versus Direct Estimates**	**Ave Error Ratio of Composite versus Direct Estimates**
Arthritis				
Dental expenditure	1	.798	.726	.748
Dental expenditure	2	.535	.684	.709
Any expenditure	1	.621	.675	.737
Any expenditure	2	.564	.560	.584
Diabetes				
Dental expenditure	1	.599	.573	.545
Dental expenditure	2	.720	.585	.651
Any expenditure	1	.670	.802	.773
Any expenditure	2	.381	.495	.609
Obesity				
Dental expenditure	1	.456	.566	.693
Dental expenditure	2	.537	.498	.496
Any expenditure	1	.688	.653	.715
Any expenditure	2	.723	.667	.623

Conclusions and Recommendations

Use of multiple years of data can decrease the standard error of a mean or proportion. If one assumes equal sample sizes and independence across years, the ratio of the standard errors of direct estimators using two and three years of data could be .71 or .58 of the standard error for a similar estimate using one year of data. However, due to use of the same set of PSU's and overlap between about half the persons in the sample in two consecutive years of data, MEPS-HC data are not independent across years and are positively correlated. This correlation lowers the reduction in the standard errors for direct estimates using multiple years of data versus the errors obtained for similar estimates using a single year of data. For highly correlated data, such as, whether a person has a certain chronic illness, we found that the errors using two and three years of data were on average .84 and .70 of the errors for similar estimates made using one year of data. For less correlated data, such as, average expenditures errors for estimates using 2 and 3 years of data were on average .77 and .65 of the errors using a single year of data.

The use of a composite estimation methodology produced reductions in errors compared to the errors of direct estimates. This was in spite of the additional errors caused by bias and parameter estimation that were included in the errors of the composite estimates. These reductions were consistent whether the compositing methodology was applied to estimates made with 1, 2, or 3 years of data. The only factor that seemed to influence the reduction gained, was the size of the states estimated. Reductions were higher for smaller states.

Although, the gain in precision of direct estimates made using multiple years of data is less than optimal due to the correlation in MEPS data across years, the gains can still be substantial. Further, the largest gains are for estimates of expenditures which tend to have the highest relative standard errors (Sommers, 2005). Using multiple years of data combined with the small area methodology can yield impressive reductions of over 50% in errors compared to the errors obtained using direct estimates and a single year of data. As with direct estimates, the best gains come for some of the worst estimates, in this case, estimates for smaller states.

The downside of using multiple years of data is interpreting what a multiple year average represents. It may be possible to calculate estimates using multiple years of data which are controlled so that when they are combined they produce the regional or national estimates for the year of interest. Future research should be performed to assess such estimates and compare their results to those of single year estimates.

References

Census Bureau Website:
http://www.census.gov/prod/2005pubs/p60-229.pdf

Cohen SB. Sample Design of the 1997 Medical Expenditure Panel Survey, Household Component. MEPS Methodology Report No 11. AHRQ Pub. No. 01-0001. Rockville, MD: Agency for Healthcare Research and Quality. 2000. http://www.meps.ahrq.gov/papers/mr11_01-0001/mr11.pdf

Monheit, A. Persistence in health expenditures in the short term: Prevalence and consequences. Medical Care, July 2003; 41(7): III-53-III-64.

Sommers, J. P. *Producing State Estimates with the Medical Expenditure Panel Survey, Household Component.* Methodology Report No. 16. December 2005. Agency for Healthcare Research and Quality, Rockville, Md.
http://meps.ahrq.gov/mepsweb/data_files/publications/mr16/mr16.pdf