EXAMPLES OF LONGITUDINAL ANALYSES USING THE MEDICAL EXPENDITURE PANEL SURVEY (MEPS)

Janet C. Greenblatt, AHRQ, 2101 E. Jefferson Street Suite 500W, Rockville, MD 20852

KEY WORDS: MEPS, NHIS, LONGITUDINAL SURVEY, SUDAAN, GEE

1. Introduction

The MEPS Household Component (HC) is an ongoing annual panel survey sponsored by the Agency for Healthcare Research and Quality. A new panel is introduced each year and data from a panel are collected over a 30-month period to obtain information that covers two consecutive calendar years. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for. MEPS also collects data on the cost, scope, and breadth of private health insurance held by and available to the U.S. population. The design of the MEPS survey permits both person based and family level estimates.

The MEPS HC sample design (1) is a stratified multistage area probability design with disproportionate sampling to facilitate the selection of oversamples of subpopulations of interest. A new panel is selected each year and followed over a 2-year period consisting of 5 data collection times. The set of households selected for the 1997 MEPS was a subsample of those participating in the 1996 National Health Interview Survey (NHIS). The NHIS is an ongoing annual household survey of approximately 42,000 households (109,000 individuals). For the Panel 2 MEPS 6,281 households were fielded (14,505 individuals) (2). The overall 1997 MEPS response rate, including responses to both the NHIS interview and the MEPS interviews, was 65.5 percent. All interviews were conducted in person, using a computer-assisted personal interview (CAPI) as the principal data collection mode.

The purpose of this paper is to describe the creation of a longitudinal file including those individuals in the panel introduced into the MEPS in 1997 (Panel 2). Two statistical tests are discussed that can be used to analyze differences and correlations among variables of interest on the MEPS longitudinal files.

2. LINK TO THE 1996 NHIS

A nationally representative subsample of 6,281 occupied dwelling units responding to the 1996 NHIS was selected for the MEPS Panel 2 introduced in January of 1997. The NHIS sample design has three stages of sample selection: an area sample of PSUs; a sample of segments (single or groups of blocks or block equivalents) within sampled PSUs; and a sample of housing units within segments. Among initially sampled households, those containing Hispanics and blacks were oversampled at rates of approximately two and 1.5 times the rate of remaining households. These same rates of oversampling are reflected in the MEPS sample of households. The only major difference in the definition of a household between NHIS and MEPS is that college aged students living away from home during the school year were interviewed at their place of residence for the NHIS but were identified by and linked to their parents' household for MEPS.

3. Construction of the Longitudinal File

A longitudinal file was created from Panel 2 individuals who responded to the survey at any time in 1997 or 1998. This file contains 12 ,445 persons, of whom 12,057 are available for longitudinal analysis over a 2-year period. There are 185 people who provided data only in 1997. These are people who participated in the survey in 1997 and died, went into a nursing home, or became ineligible for the survey for some other reason, such as they entered into the military or left the country later in 1997. In addition, there are 203 people who provided data only in 1998. These are newborns or those who came into a selected household for the first time in 1998, such as those who moved into a sample household from a nursing home or other institution. The sample that exists for only 1 year is provided on the file to facilitate analyses that cover the experience of the U.S. civilian non-institutionalized population over 1997 and 1998. The resulting file contains five variables: DUPERSID, VARSTR, VARPSU, YEARIND, and LONGWGT. DUPERSID identifies respondents to the survey, VARSTR and VARPSU define the sample strata and PSUs for calculation of variances. YEARIND is used to distinguish individuals who appeared in both 1997 and 1998, in only 1997 and in only 1998. The variable LONGWGT is an adjusted weight variable that can be used to produce average annual estimates for

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

respondents to both 1997 and 1998 or separate annual estimates for 1997 or 1998.

Codebook

VARIABLE	DESCRIPTION
DUPERSID	Sample Person ID
LONGWGT	Panel 2 Longitudinal weight nonresponse and poststratified to CPS totals
VARSTR	Panel 2 Stratum and PSU for either 1997 or 1998 variance calculation
VARPSU	
YEARIND	Panel 2 Year Indicator 1=in Both 1997 and 1998 2=in only 1997 3=in only 1998

The MEPS website maintained by AHRQ http://www.meps.ahrq.gov currently contains three longitudinal files: the Panel 1 file, introduced in 1996, the Panel 2 file, introduced in 1997, and the Panel 3 file, introduced in 1998. To obtain analytic variables, the records on this file need to be linked to the 1997 and 1998 MEPS public use data sets by the sample person identifier (DUPERSID). The 1997-1998 panel can also be linked to 1996 NHIS data, thereby providing three years of longitudinal data on Panel 2.

4. Analysis of Longitudinal MEPS Data

To obtain estimates of variability (such as the standard error of sample estimates or corresponding confidence intervals) for estimates based on MEPS survey data, the complex sample design of MEPS for both person and family level analyses must be taken into account. Various approaches can be used to develop such estimates of variance including use of the Taylor series or replication methodologies. Replicate weights have not yet been developed for the MEPS 1997 data. Using a Taylor Series approach, variance estimation strata and the variance estimation PSU's within these strata must be specified. The corresponding variables on the 1997 MEPS full year utilization database are VARSTR97 and VARPSU97, respectively. Specifying a "with replacement" design in a computer software package, such as SUDAAN should provide standard errors appropriate for assessing the variability of MEPS survey estimates. Two methods of analyzing longitudinal data will be discussed below: Estimating Transitional Probabilities and fitting Logistic Regression Models for Clustered Data.

5. Transitional Probabilities

An example of a study question that might be asked under this type of analysis is "Does the probability of becoming limited in activities increase with age, after controlling for gender and number of office-based doctor visits in 1998." For this example, the file was limited to those persons who responded to the survey in both 1997 and 1998. Those records with YEARIND=1 (in both years) were extracted from the 1997 Panel 2 Longitudinal File. The identification numbers (DUPERSID) of the persons in both years of the survey were merged to the 1997 and 1998 MEPS public use files and analysis variables were extracted and merged to the longitudinal records. The resulting file contained one record per person with 1997 and 1998 variables included.

The analysis was restricted to those persons who said they had no limitation in activities in 1997. They were subsequently coded as '1' if they developed a limitation in 1998 and '0' if they did not develop a limitation in 1998. The variable called "LIMITED" was constructed as follows: Individuals were asked at each interview time if they had any difficulty, because of a health or physical problem in performing certain specific physical actions such as walking, climbing stairs, grasping objects, reaching overhead, lifting, bending or stooping, or standing for long periods of time. They were also asked if they had any limitation in work, housework, or school or if they experienced confusion or memory loss or had difficulties making decisions.

The SUDAAN program PROC RLOGIST was run for this dataset with the following code:

```
PROC RLOGIST
nest VARSTR VARPSU;
weight LONGWGT;
model LIMITED=AGECAT SEX
          OFFBASED98;
run;
```

The output from PROC RLOGIST consists of Wald Chi Squares and P-values, odds ratios, and upper and lower 95% boundaries of the odds ratio. An example of the output for the variable AGECAT is shown below:

```
Transitional Probability Output
Variable  Odds Ratio  Lower 95% Upper 95%
AGECAT
 18-25     1.00        1.00       1.00
```

25-44	1.65	1.09	2.52
45-64	3.82	2.63	5.54
65+	7.57	4.91	11.67

The results of this analysis show that, after controlling for sex and number of office-based physician visits, the probability of becoming limited in activities from 1997 to 1998 increases with age. Persons age 65 and older were nearly 8 times more likely than those age 18-25 and about twice as likely as those age 45-64 to become limited in activities.

6. Regression for Clustered Data

A study question appropriate for this type of analysis is "Does the probability of having office-based doctor visits increase from 1997 to 1998, after controlling for insurance coverage, age and sex?" This type of analysis takes into account the fact that we have observations for year 1 and year 2 from the same person and therefore, these observations are correlated. This method fits transition models using generalized estimating equations (GEE). The analysis population for this example was all adults age 18 and older. The dependent variable was the number of office-based doctor visits, and the independent variables were age category, poverty status, and race/ethnicity.

An analytic file was created, that extracted YEARIND=1 (in both years of survey) from the Longitudinal File. The DUPERSID identification numbers were matched to the 1997 and 1998 Public Use Files to extract variables of interest. The analytic file contained 2 records per person, one with 1997 and another with 1998 data. A new variable was created called 'YEAR' which was '1' for the 1997 record and '2' for the 1998 record. Because the dependent variable was continuous, SUDAAN, PROC REGRESS, was used with options for clustered data as follows:

```
PROC REGRESS, r=EXCHANGEABLE, steps=3,
    semethod=ZEGER;
nest VARSTR VARPSU CASE;
model OFFVISIT=YEAR AGECAT POVSTAT
    RACE;
run;
```

where:
r=EXCHANGEABLE speciifies equal pairwise correlations
rsteps=3 specifies the maximum number of steps used to fit the model
semethod=ZEGER specifies full robust or sandwich variance estimator.

The output from the linear regression shows the p-values associated with the correlation matrix and the Wald F statistic for the variable "YEAR". If the p-value is significant at the .05 level, then the analyst can be 95% certain that there was a statistically significant change in number of office based doctor visits between 1997 and 1998, after controlling for age, poverty status, and race/ethnicity. The output from PROC REGRESS is shown below:

Contrast	Wald F	P-value Wald F
OVERALL MODEL	172.60	0.0000
MODEL MINUS INTERCEPT	55.46	0.0000
Intercept	.	.
YEAR	0.80	0.3721
AGECAT	90.67	0.0000
SEX		
RACETHNX	133.26	0.0000
POVERTY STATUS	4.36	0.0019

The output from this linear regression shows that the p-value for the variable "YEAR" is not significant at the .05 level indicating that there was no change in number of office based doctor visits between 1997 and 1998, after controlling for age, sex, poverty status, and race/ethnicity.

7. CONCLUSIONS

A new panel is introduced into MEPS every year and followed for a 2-year period. Using the NHIS file of respondents from which the MEPS panel sample is drawn, an analyst could construct a file with 3-years of data for selected variables. These data could provide a valuable source for analyzing trends over time. There are many methods for analyzing longitudinal data, two of which were discussed here: Transitional Probability Analysis and Regression for Clustered Data. The selection of a method for analysis is dependent on the study hypothesis. Whatever method one uses, the complex survey design of the MEPS must be taken into account through the use of programs that encorporate the survey design into the analysis such as SUDAAN, STATA, and WESVAR.

8. References

1. Cohen SB, Sample design of the 1997 Medical Expenditure Panel Survey Household Component. Rockville (MD): Agency for Health Care Research and Quality: 2000. MEPS Methodology Report No. 11. AHRQ Pub. No. 01-0001.

2. MEPS HC-022: 2000 P4R3/P5/R1 Population Characteristics. Rockville (MD): Agency for Healthcare Research and Quality. AHRQ Clearinghouse Number 01-DP06.

3. MEPS HC-020: 1997 Full Year Consolidated File. Can be downloaded from http://www.meps.ahrq.gov/Puf/PufDetail.asp

4. MEPS HC-028: 1998 Full Year Consolidated File. Can be downloaded from http://www.meps.ahrq.gov/Puf/PufDetail.asp

5. Diggle PJ, Liang KY and Zeger SL. (1994) Analysis of Longitudinal Data. NY: Oxford University Press Inc.

6. Korn, EL and Graubard, BI. 1999. Analysis of Health Surveys. John Wiley & Sons, Inc.

7. Shah BV, Barnwell BG, and Bieler GS (1997). SUDAAN User's Manual, Release 7.5. Research Triangle Park, NC: Research Triangle Institute.

8. Zeger S. and Liang K. (1986). *Longitudinal data analysis for discrete and continuous outcomes. Biometrics* 42, 121-130.

9. Liang K. and Zeger S. (1986). *Longitudinal data analysis using generalized linear models*. *Biometrika* 73, 13-22.