Class Variables for MEPS Expenditure Imputations

Marc W. Zodet and Diana Z. Wobus and Steven R. Machlin and David Kashihara and Deborah D. Dougherty

Agency for Healthcare Research and Quality Working Paper No. 04005

October 2004

Suggested citation: Zodet MW, Wobus DZ, Machlin SR, Kashihara D, Dougherty DD. Class Variables for MEPS Expenditure Imputations. Agency for Healthcare Research and Quality Working Paper No. 04005, October 2004, http://www.ahrq.gov.

AHRQ Working Papers provide preliminary analysis of substantive, technical, and methodological issues. The papers are works in progress and have not undergone a formal 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.

Class Variables for MEPS Expenditure Imputations
Marc W. Zodet, Diana Z. Wobus, Steven R. Machlin, David Kashihara, and Deborah D. Dougherty
October 2004

ABSTRACT

The Medical Expenditure Panel Survey (MEPS) collects data on healthcare utilization, expenditures, sources of payment, insurance coverage, and healthcare quality measures. The survey was designed to produce national and regional estimates for the U.S. civilian non-institutionalized population. The data on medical expenses are collected from both household respondents in the Household Component and from a sample of their health care providers in the Medical Provider Component. In the absence of payment information from either component, expenditure data are derived for sample persons through an imputation process. Missing expense data are imputed at the event level for each medical event type using a weighted hot-deck procedure. This process utilizes individual and event level data collected in MEPS that are correlated with medical expenditures. Bivariate analyses and linear regression models were utilized to assess the current class variables used for imputation. This paper details the methodology used to select, prioritize, and categorize the class variables used to impute missing expenditures for three event types: doctor visits, hospitalizations, and home health visits.

Marc Zodet
Statistician, Center for Financing, Access, and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
E-mail: MZodet@ahrq.gov

Diana Z. Wobus and Deborah D. Dougherty
Westat
Research Boulevard
Rockville, MD 20850

Steven R. Machlin and David Kashihara
Center for Financing, Access, and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

Introduction

The Medical Expenditure Panel Survey (MEPS) collects data on health care utilization, expenditures, sources of payment, insurance coverage, and health care quality measures. The survey, conducted annually since 1996 by the Agency for Healthcare Research and Quality (AHRQ), is designed to produce national and regional estimates for the U.S. civilian non-institutionalized population.

MEPS data on medical expenses are collected from both household respondents in the Household Component and from a sample of their health care providers in the Medical Provider Component. When payment (i.e., expenditure) information is missing from either component, these data are derived for sample persons through an imputation process. Expense data are collected at the event level for each medical event type and a weighted hot-deck procedure is used for imputation. This process utilizes individual and event level data collected in MEPS that are correlated with medical expenditures. AHRQ uses bivariate analyses and linear regression models to assess potential variables use in imputation.

Using office-based visits and inpatients stays as examples, this paper details the methodology used to select, prioritize, and categorize the class variables used to impute missing expenditure data. The paper does not address the specifics of how the imputations are actually carried out. For a more detailed description of the imputation procedure see Machlin and Dougherty (2004).

Background

Class Variables

A key component of a hot-deck procedure is the matching of sample observations with missing information (i.e., recipients) to similar sample observations not missing the information (i.e., donors). Categorical or “class” variables that characterize the sample observations are used to classify both recipients and donors into imputation cells or classes. Within each imputation cell, the recipients’ missing values are imputed from the values of the donors. Variables that are considered important predictors of the data to be imputed are the primary candidates for use as class variables. The underlying assumption is that the recipients have similar values with regard to the measure of interest as the donors and that the data associated with the donors within the same imputation cell are appropriate for the imputation of the missing values (Cox, 1980).

Class variables are typically ordered in accordance with predictive importance (i.e., more important predictors ranked higher). If there are fewer donors than recipients in a cell, then the procedure will begin collapsing over the categories of the class variables, starting at the bottom of the list and working up, until a sufficient number of donors are available.

MEPS Event Types

MEPS expenditure data are imputed separately for each of twelve event types: hospital inpatient stays, outpatient, emergency room, office-based visits (physician and non-physician), home health (agency and paid independent), dental, other medical equipment/supplies, and prescription medications. Separate imputations are conducted for each event type because the relevant variables and statistically significant correlates are not consistent across the event types. Therefore, for each event type, the class variables are evaluated and chosen separately, but some of the same class variables are used across different event types. For example, the class variables for the imputations of both emergency room expenditures and dental expenditures include patient age. While the same class variable may be used across multiple event types, the specification of the specific categories for the variable used in the individual imputations may differ. Regression methods are used to support the selection of the majority of class variables. The remainder of this paper discusses the process by which variables are evaluated and selected for use in the creation of imputation cells.

Methodology

The lists of class variables used to impute event-specific expenditures were initially established based on the first year of MEPS data (1996). The process of identifying predictors of total expenditures was based both on substantive decisions as well as statistical associations, that were identified primarily through multiple linear regression models. In 2002, analysts from AHRQ and Westat, the data collection contractor, jointly began to reevaluate and revise these lists of class variables. The methods presented in this section and the Examples section below are reflective of those efforts and focus primarily on the quantitative methods used in the decision process.

Data

Data for this project came from the MEPS event level files. Only events that were potential donors (i.e., complete on the household component and/or the medical provider component) were used in the analyses. Multiple years of data were examined: 1997, 1998, and 1999. For the most part, each year of data was examined separately. However, when the numbers of events were small (e.g., home health services) years of data were pooled to stabilize the variance of the estimates.

Potential Class Variables

The class variables considered for the imputation were those collected in MEPS that were thought a priori to potentially have a significant impact on total expenditures. Two variables were considered important enough to be included in all imputation procedures: type of insurance coverage and total charges. The former was chosen because the generosity of payment for health care services can vary widely depending on the type of insurance one carries (e.g., Uninsured, Private, Medicare, Medicaid, etc.). The later was chosen because total charges are highly correlated with total expenditures. Unfortunately, when expenditures are missing total charges are also frequently missing.

Other predictors of expenditures were selected quantitatively. These included various indicators of health care services (e.g., laboratory tests, radiology, surgeries/extractions, etc.). Predictors can be specific to the type of event. For example, the number of nights is associated with inpatient hospital stays, but was not relevant to physician office visits.

Regression Models

Multiple linear regression was used to evaluate the statistical associations between potential class variables and total expenditures. The dependent variable in each model was total expenditures for the event. Total expenditures were defined as the sum of direct payments for care provided during the year, including both out-of-pocket and third-party (e.g., private insurance, Medicare, and Medicaid) payments.

Two approaches were taken when fitting the regression models to assess the association between potential class variables and total expenditures. First, to adjust for the complex design of MEPS, linear regression models were fit using PROC REGRESS in the SUDAAN statistical software package (www.rti.org). With these models, the two primary considerations were: 1) whether or not the resulting regression coefficients were significant and 2) the relative magnitude and direction of the significant coefficients. Statistical significance was determined at the α=0.05 level. To provide additional guidance in the selection of variables, models were fit using SAS PROC STEPWISE (www.sas.com) The significance level for entry and retention was 0.15 (the SAS default). Block entry grouping of variables was used to ensure that all levels of a particular variable were entered/retained as a group.

Results from both sets of models (i.e., those fit using SUDAAN and those fit using SAS) were considered when selecting the final list of class variables to be used in the imputation procedures. Model results were also used to prioritize the class variables, which were ranked with the most important substantive and statistical predictors placed higher on the list. Model results were also used to determine the collapsing strategies for variables with three or more levels. When it become necessary to collapse over imputation cells due to insufficient availability of donors, the most important predictors of total expenditures (i.e., those higher on the list) were preserved. This was an effort to assure that recipients and donors were matched based on the most important predictors of total expenditures.

Examples

As noted previously, the process for identifying class variables was performed separately for each type of event. Examples of how this process works for physician office visits and inpatient hospital stays are presented below. To provide a point of reference for the magnitude of total expenses attributed to each of these two types of medical events, Table 1 presents mean total expenditures per event for 1997 through 1999 for events with complete (i.e., not imputed) data. In 2001, approximately one-third of the expenditure values were fully imputed for physician office visits and hospital inpatient stays.

Table 1. Total Expenditures for Physician Office Visits and Inpatient Hospital Stays by Year, Mean (Std Err)

	1997	1998	1999
Physician Office Visits[1]	$92 ($3)	$98 ($3)	$107 ($3)
Hospital Inpatient Stays[1,2]	$5,647 ($301)	$5,375 ($304)	$5,929 ($367)

[1] Estimates are for patients with complete event data (i.e., donors).
[2] Only events of patients who did not die during the year.

During the late 1990's, total expenditures for a physician office visit averaged roughly $100 per event while facility expenditures for an inpatient hospital stay during this same period averaged approximately $5,600 per event.

Physician Office Visits

Table 2 summarizes regression models fit using SUDAAN (i.e., adjusted for the complex survey design). Separate models were fit for the years 1997, 1998, and 1999 with physician office visit expenditures as the dependent variable in each model. Independent variables in the models were those hypothesized as potentially significant predictors of office visit expenditures and were the candidate variables from which to select the class variables to create the imputation cells.

The information provided in Table 2 shows that surgery, radiology, other services, and laboratory services were all statistically significant predictors of physician office visit expenditures across all three years (p-values < 0.01). Other variables were statistically significant predictors in some years, but not others. For example, patient age was highly significant (p-value < 0.01) in 1999, but not in the two preceding years.

Results from fitting the STEPWISE models for each year are presented in Table 3, which shows the order in which the independent variables entered into the models. Surgery, radiology, and other services were consistently the first, second, and third variables entered into the model each year. Perceived health and laboratory services alternated as the fourth and fifth variables depending on the year.

Table 2. P-Values (Wald F Statistics) from Weighted Regression Models by Year (SUDAAN), Dependent Variable = Physician Office Visit Expenditures.

	1997	1998	1999
# Obs Used in Regression	48,815	34,948	31,978
R^2	0.043	0.048	0.032
Class Variable[1]			
Surgery (Yes; No)	<0.01	<0.01	<0.01
Radiology (Yes; No)	<0.01	<0.01	<0.01
Other Services (Yes; No)	<0.01	<0.01	<0.01
Laboratory Services (Yes; No)	<0.01	<0.01	<0.01
Saw Non-MD (Yes; No)		<0.10	<0.10
Age (<18; 18-24; 25-64; 65+)			<0.01
Perceived Health (Poor; Other)		<0.10	<0.05
Race/Ethnicity (Hispanic; Other)			
Census Region (S; MW; NE; W)			
MSA (MSA; Non-MSA)	<0.05	<0.10	

[1] Variables forced into the models are not shown (e.g., Insurance Source of Payment (Private; Medicare; Medicaid; CHAMPUS/TRICARE), Decile of Total Charges, and HMO Indicator (Yes; No))

Table 4 presents the SUDAAN regression coefficients for selected variables used in the model. This table illustrates that surgery was consistently associated with higher physician office visit expenditures. For the years observed (i.e., 1997-1999), the average additional expenditures associated with having a surgical procedure during a physician office visit was approximately $200 when controlling for the other variables on the model. These additional expenditures were substantially greater than what is observed for the other factors being considered. For example, the difference in mean expenditures per event associated with surgery compared to radiology ranged from approximately $115 (1999) to approximately $136 (1997).

Table 3. Order of Entry into Weighted Regression Models by Year (STEPWISE Procedure), Dependent Variable = Physician Office Visit Expenditures.

	1997	1998	1999
# Obs Used in Regression	48,815	34,948	31,978
R^2	0.042	0.048	0.032
Variable Entry Order			
1st	Surgery	Surgery	Surgery
2nd	Radiology	Radiology	Radiology
3rd	Other Services	Other Services	Other Services
4th	Prcvd Health	Lab Services	Prcvd Health
5th	Lab Services	Prcvd Health	Lab Services
6th	Saw NonMD	Age	Age
7th	Region	Saw NonMD	Region
8th	Region	Region	Saw NonMD

Among the other four highly significant variables (i.e., surgery, radiology, other services, and laboratory services) the magnitudes of the coefficients (i.e., the average expenditures) associated with a particular service tended to diminish in accordance with the entry order of the variables into the STEPWISE models. However, while the expenses associated with surgery were consistently higher than those of any of the other factors considered, the magnitude of the differences between the other services (i.e., radiology, other, and laboratory) varied from year to year. For example, a simple comparison of the mean office visit expenditures associated with radiology compared to other services demonstrated no difference in 1997; but there was a significant difference in 1998, with payments for office visits involving a radiology service running about $35 more per visit compared to those with other services. In summary, of the factors considered, surgery clearly had the greatest impact on increasing physician office visit expenditures.

Table 4. Coefficients for Select Variables from Weighted Regression Models by Year (SUDAAN), Dependent Variable = Physician Office Visit Expenditures.

	β-Coefficients (SE β-Coefficients)		
	1997	1998	1999
Class Variable			
Surgery	$205 ($25)	$198 ($28)	$196 ($28)
Radiology	$69 ($5)	$79 ($7)	$81 ($9)
Other Services	$53 ($9)	$44 ($10)	$58 ($8)
Lab Services	$21 ($4)	$24 ($6)	$20 ($6)
Perceived Health	$40 ($25)	$30 ($17)	$34 ($14)
Saw Non-MD	-$8 ($6)	-$14 ($8)	-$11 ($7)

Table 5. Final Class Variable List for Imputing Physician Office Visit Expenditures.

1. HMO
2. Type of Insurance Coverage
3. Total Charges
4. Surgery
5. Radiology
6. Other Services
7. Laboratory Services
8. Perceived Health
9. Saw Non-MD

The final list of class variables used to impute office visit expenditures is presented in Table 5. The top three variables were chosen based upon substantive reasoning: HMO (an indicator of whether or not the patient was enrolled in an HMO), type of insurance coverage, and total charges. The remainder were chosen based upon the regression results. Surgery, radiology, and other services followed in that order primarily because they were each highly significant in each of the SUDAAN models across all three years and because they were consistently the first three variables entered into the STEPWISE models in all three years. The laboratory services variable was placed above the perceived health variable because it was more highly significant in each of the SUDAAN models and because it entered into the STEPWISE models before the perceived health variable for two of the three years. In turn, the perceived health variable was more statistically significant in the SUDAAN models than the saw non-MD variable.

It also entered into the each of the STEPWISE models before saw non-MD and was therefore higher on the list. Despite being statistically significant in at least one of the years examined, neither age nor MSA were included on the final list of class variables. The rationale for dropping age and MSA came from the fact that age was only significant in one year (p-value < 0.01) and MSA was never retained in any of the STEPWISE procedures.

Hospital Inpatient Stay

Table 6 shows that, based on the SUDAAN model, the only statistically significant predictors of inpatient hospital stay expenditures of the variables considered were length of stay and reason in hospital (p-values < 0.01). These results were consistent across each of the three years. Results from the STEPWISE models confirmed the importance of both length of stay (LOS) and reason in hospital as these variables were consistently the first and second variable respectively added to each of the models (Table 7).

Table 6. P-Values (Wald F Statistics) from Weighted Regression Models by Year (SUDAAN), Dependent Variable = Inpatient Hospital Stay Expenditures.

	1997	1998	1999
# Obs Used in Regression	1,881	1,294	1,259
R^2	0.40	0.36	0.44
Class Variable[1]			
ER before Admission (Yes; No)			
HMO (Yes; No)			
Length of Stay (0, 1, 2,…6, 7, 8-13, 14-30, 31-60, 61+)	<0.01	<0.01	<0.01
Reason in Hospital (Surgery; Treatment/Therapy; Diagnostic Tests; Give Birth; To be Born; Other)	<0.01	<0.01	<0.01
Census Region (N; MW; S; W)			
MSA (MSA; Non-MSA)			

[1] Variables forced into the models are not shown (e.g., Insurance Source of Payment (Private; Medicare; Medicaid; CHAMPUS/TRICARE) and Decile of Total Charges)

Table 7. Order of Entry into Weighted Regression Models by Year (STEPWISE Procedure), Dependent Variable = Inpatient Hospital Stay Expenditures.

	1997	1998	1999
# Obs Used in Regression	1,881	1,294	1,259
R^2	0.32	0.31	0.31
Variable Entry Order			
1st	LOS	LOS	LOS
2nd	Reason	Reason	Reason
3rd	ER before	Region	Region
4th	Region		HMO

The coefficients for length of stay and reason in hospital that resulted from the design adjusted regressions are presented in Table 8. For the most part, mean expenditures per stay increased as the length of stay increased. There was some erratic behavior of the coefficients for the longest lengths of stay (e.g., sharp drops in average expenditures associated with lengths of stay of more than sixty days). While this may have been due to the influence of outliers and/or may suggest some other functional form of the variable was more appropriate, it had no impact on our decision to include length of stay as a high priority variable. Surgery was the most significant contributor to inpatient expenditures compared to the other reasons for hospitalization. The coefficients indicated that surgery is associated with an approximate increase in inpatient expenditures of at least $3,000 compared to the other reasons for admission to the hospital.

Table 8. Coefficients for Select Variables; Weighted Regression Models by Year (SUDAAN), Dependent Variable = Inpatient Hospital Stay Expenditures.

		β-Coefficients (SE β-Coefficients)		
		1997	1998	1999
Class Variable				
Length of Stay (days)	0 (Reference)	$0 ($0)	$0 ($0)	$0 ($0)
	1	$2,121 ($411)	$2,020 ($488)	$771 ($550)
	2	$3,824 ($448)	$3,073 (480)	$2,146 ($638)
	3	$4,715 ($523)	$3,792 ($505)	$3,126 ($569)
	4	$5,637 ($615)	$5,239 ($727)	$4,193 ($708)
	5	$6,922 ($933)	$6,624 ($976)	$4,436 ($707)
	6	$7,853 ($836)	$7,307 ($1,236)	$6,165 ($1,125)
	7	$8,532 ($927)	$7,180 ($1,110)	$7,340 ($1,066)
	8-13	$10,555 ($1,053)	$8,722 ($761)	$8,769 ($1,124)
	14-30	$18,967 ($3,048)	$18,123 ($2,706)	$19,409 ($4,170)
	31-60	$44,950 ($12,311)	$25,739 ($6,567)	$39,188 ($17,209)
	61+	$5,484 ($827)	$15,107 ($9,416)	$48,210 ($11,756)
Reason in Hospital	Surgery (Reference)	$0 ($0)	$0 ($0)	$0 ($0)
	Treatment / Therapy	-$4,342 ($590)	-$3,906 ($676)	-$4,937 ($882)
	Diagnostic Tests	-$4,315 ($570)	-$3,543 ($521)	-$4,998 ($734)
	Give Birth	-$3,380 ($461)	-$3,122 ($532)	-$3,780 ($622)
	To be Born	$2,456 ($4,525)	-$2,082 ($1,956)	-$6,554 ($1,701)
	Other	-$3,792 ($924)	-$3,600 ($1,525)	-$4,567 ($796)

The final list of class variables used to impute inpatient hospital expenditures is presented in Table 9. As usual, type of insurance coverage and total charges were included at the top of the list. In addition, an indicator of whether or not there was an emergency room event before the hospital admission was included because the billing information for the ER and the hospital stay are often rolled up into one expenditure figure for the stay. Based on the findings noted above, length of stay and reason in hospital then followed in that order. MSA status and census region were also included on the final list; based in part, on their being retained in the STEPWISE models (p-values<0.15).

Table 9. Final Class Variable List for Imputing Inpatient Hospital Expenditures.

1. Type of Insurance Coverage
2. Total Charges
3. ER before Admission
4. Length of Stay
5. Reason in Hospital
6. MSA/Non-MSA
7. Census Region

Class Variable Collapsing Strategy

Results from the regression modeling presented above were also used to establish the collapsing strategy used during the hot-deck procedure for variables with three or more levels. The coefficients from the design adjusted regression models weighed heavily in deciding how to collapse over variables with three or more categories. For example, consider the reason in hospital variable described above. Note that there was little difference between the coefficients for *treatment/therapy* and *diagnostics tests only*. Hence, prior to using the variable in the imputation procedure it seemed reasonable to recode these two levels into one; effectively reducing the variable from six levels to five levels (Table 10). During the imputation procedure, further collapsing of the remaining levels was determined by the number of recipients/donors residing in a given imputation cell. Given the findings noted above, it was important to maintain surgery as a separate category whenever possible since it was associated with the highest mean expenditures. Thus, the hot-deck was programmed to maintain surgery as a separate category whenever possible.

Summary

The process of selecting the most appropriate class variables to use when imputing health care expenditures is a combination of art and science that involves both substantive reasoning and statistical analysis. As illustrated above, predictors of expenses can vary by event type and the selection of class variables includes the examination of both person characteristics and event characteristics. Careful selection of class variables should improve the quality of the hot-deck imputation procedure and reduce bias in MEPS expenditure estimates. The class variables used to impute health care expenditure data in MEPS are periodically reviewed and refined. Class variables being considered for future inclusion in the imputation procedures include provider specialty for ambulatory events and person-level condition information.

Table 10. Coefficients for *Reason in Hospital*; Weighted Regression Models by Year (SUDAAN), Dependent Variable = Inpatient Hospital Stay Expenditures.

			β-Coefficients (SE β-Coefficients)		
			1997	1998	1999
Reason in Hospital Recoded into a single category (i.e., Reason in Hospital changes from 6-level variable to a 5-level variable)	{	Surgery (Reference)	$0	$0	$0
		Treatment / Therapy	-$4,342 ($590)	-$3,906 ($676)	-$4,937 ($882)
		Diagnostic Tests Only	-$4,315 ($570)	-$3,543 ($521)	-$4,998 ($734)
		Give Birth	-$3,381 ($461)	-$3,122 ($532)	-$3,780 ($622)
		To be Born	$2,456 ($4,525)	-$2,082 ($1,956)	-$6,554 ($1,701)
		Other	-$3,792 ($924)	-$3,600 ($1,525)	-$4,567 ($796)

References

Cox B. (1980). The Weighted Sequential Hot Deck Imputation Procedure. *American Statistical Association 2004 Proceedings of the Section on Survey Research Methods*, 721-726.

Machlin S. and Dougherty D. (2004). Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey. *American Statistical Association 2004 Proceedings of the Section on Survey Research Methods*.