Evaluation of Health Care Event Reporting in a National Household Survey

Jerrod Anderson[1], Emily Mitchell[1], Adam Biener[1]
[1]Division of Statistical Research and Methods, Center for Financing, Access and Cost Trends, Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20852

Abstract
The Medical Expenditure Panel Survey (MEPS) is a nationally representative health survey conducted annually by the Agency for Healthcare Research and Quality (AHRQ). Respondents to the Household Component (HC) of MEPS provide detailed information on health care events in addition to socioeconomic data. For a subset of respondents, medical providers that are associated with health events reported by the household are contacted to obtain more precise information on event details and expenditures. While the primary motivation for conducting this follow-back survey, called the Medical Provider Component (MPC), is to collect data for improving the quality and completeness of expenditure data for household-reported events, we leverage MPC information to determine the extent to which HC respondents may be mis-reporting the number of medical events for sample persons. We treat MPC data as a validation data set for household responses and use machine learning methods to identify characteristics of reporting accuracy and to predict reporting accuracy.

Key Words: *Data quality, survey accuracy, health data, medical events, MEPS, machine learning*

1. Introduction

The Agency for Healthcare Research and Quality's Medical Expenditure Panel Survey (MEPS) is a nationally representative survey of the U.S. civilian non-institutionalized population that has been conducted annually since 1996. For this study we focus on two components of MEPS: the Household Component (HC) and the Medical Provider Component (MPC). In the HC, each household (usually with one respondent per household) is interviewed five times over the course of 2.5 years, to gather health status and expenditure information covering two calendar years. The MPC surveys the medical providers from a subset of the persons in the HC in order to obtain more detailed information about expenditures and sources of payment for specific types of health events reported in the survey (Machlin and Dougherty (2007)). In this analysis we utilize the MPC as a validation data set to measure the accuracy of the HC, though because the MPC is a survey of providers the measure of accuracy will only be an approximation.

Reporting error is a concern in all large national surveys, including MEPS, as it may lead to biased estimates of healthcare use and medical expenditures. Previous research has found evidence of under-reporting among Medicare beneficiaries in the MEPS sample (Zuvekas and Olin (2009a, 2009b)), while Hill, Zuvekas, and Zodet (2011) found that overall drug fills and expenditures are measured accurately, but the number of different drugs used was under-reported and number of fills per drug was over-reported. Because respondents may not always provide accurate or complete information, the goal of this analysis is to use the MPC data to validate the responses observed in the HC and to measure the accuracy with which households report the number of events (i.e., healthcare utilization). We also use these data to construct predictive models to aid in understanding factors that contribute to inaccurate reporting, as well as to assess our ability to predict which persons are more likely to under or over report.

2. Data and Methods

The data used in this analysis come from the 2013 HC and MPC surveys. We compare the total number of reported events from each component and classify individuals as over-, under-, or equal reporters using the MPC as the accurate count. The number of events is the sum of inpatient and outpatient hospital events, emergency room events, and physician related office-based events. The study sample is restricted to persons age 18 or older who had at least one event reported that occurred in 2013. Since reporting accuracy can vary across interviews, the data are analyzed at the person-round level.

The outcome we study is whether an individual over- under- or equal-reported the total number of medical events within an interview round. Each model includes socio-demographic variables (e.g. age, race, poverty status), health status variables (e.g. health conditions, perceived health status), and interview para-data variables (e.g. interview length, experience of interviewer). The data are divided into training and test sets with 13,558 observations in the training data and 3,388 in the test data.

Two types of predictive models are used. The first is a multinomial regression model, implemented through the caret package in R. The general form of the model is,

$$\eta_{ij} = \alpha_j + \boldsymbol{\beta}_j \boldsymbol{X}_i + \boldsymbol{\gamma}_j \boldsymbol{Y}_i + \boldsymbol{\delta}_j \boldsymbol{Z}_i, \qquad \text{(Equation 1)}$$

$$\pi_{ij} = \frac{\exp\{\eta_{ij}\}}{\sum_{k=1}^{3} \exp\{\eta_{ik}\}}, \qquad \text{(Equation 2)}$$

where X_i is a set of socio-demographic variables for individual i, Y_i is a set of health status variables for individual i, Z_i is a set of survey paradata variables for individual i, and π_{ij} is the probability that individual i is in category j, where j = under-, equal-, and over-reporting (equal-reporting is the reference category).

The second predictive model is built utilizing the gradient boosting machine (GBM) algorithm as implemented in the caret package in R. GBM is a tree-based machine learning algorithm used for classification and regression. It combines multiple models algorithmically and evaluates residuals iteratively. This algorithm was chosen over the classification and regression tree (CART) algorithm and a bagged random forest algorithm on the training data because it attained the highest accuracy when tested against the training data. The predictive accuracy of the multinomial and GBM models is determined by applying the models to the test data set.

Table 1 presents the variables used for both the multinomial and GBM models.

Table 1: Independent Variables Used

Socio-demographic	Health Status	Survey Paradata
Sex (male vs. female)	Perceived mental health status	Length of interview
Age	Perceived health status	Interview language
Race/ethnicity	Ever had cancer	Length of time period for which information is collected
Poverty status	Ever had high cholesterol	Usage of memory aids
Census region	Ever had diabetes	Usage of records
Type of insurance (if any)	Ever had emphysema	Number of persons in household
Highest level of education attained	Ever had high blood pressure	Age profile of others in household
Marital status	Ever had stroke	Number of events per household
Employment status	Pregnancy status	Percent of observations that were repeat visits
		Percent of observations that were lab tests
		Percent of events for which respondent could not give specific day
		Percent duplicate (percent of visits reported on the same day for the same provider)
		Relationship to respondent
		Number of events in previous round
		Proxy reported (someone from outside the household is the respondent)
		Number of events in current round
		Experience of interviewer
		Number of household-reported events
		Number of household-reported events in previous round

3. Results

3.1 Multinomial Results

Table 2 provides summary data on the prevalence of the types of reporting behavior. The percentage of person-rounds that under-, equal-, and over-report are 44, 40, and 16 percent, respectively. Figure 1 presents the distribution of the number of events reported for each person-round, where the plurality of persons have 1 reported healthcare event per round. The reporting behavior by poverty status and educational attainment are provided in Figures 2 and 3, respectively.

Table 2: Observations by Reporting Status

	Under-report (HC < MPC)	Equal-report (HC = MPC)	Over-report (HC > MPC)
Number of person-rounds	7,496	6,752	2,698
Percent of total observations	44.2	39.8	15.9

The three independent variables with the largest coefficients (all of which are statistically significant at the 5% level) of the multinomial analysis are presented in Table 3. The coefficients show the marginal effect (on the logit) of each variable relative to the reference category of equal-reporting. For example, if the individual is the child or grandchild of the respondent, then this increases the logit (equation 1) by 1.02. This translates to a 36 percent increase in the probability of the individual being categorized as an under-reporter relative to equal reporting. Note that the predictors with the largest coefficients are paradata variables. It should also be noted that while paradata variables provide the largest coefficients, under- and over- reporting appear to be driven by different processes, suggesting that attempts to mitigate these problems will have different solutions. The multinomial model provides a relatively poor model for predicting report status. When run on a test dataset, the model has an accuracy of 59 percent and a Cohen's kappa (a measure of agreement between prediction and observed outcome that takes the probability of expected agreement into account) of 26. Because the poverty status and educational attainment are of interest to many researchers, the predicted probability of reporting status by poverty status and educational attainment are shown in Figures 4 and 5, respectively.

3.2 Gradient Boosted Models Results

In order to improve predictive accuracy, we estimated alternative models using the GBM algorithm. The top five variables of importance are total reported events, percent lab tests, percent repeat visits, percent of events with no specific day reported, and number of reported events in previous round. The GBM algorithm provides much more predictive power than the multinomial model. When applied to the test data set, GBM provides an accuracy of 78 percent with a kappa of 64. The confusion matrix from running the model on the test data is presented in Table 4. The accuracy of the GBM algorithm in predicting

Table 3: Coefficients from Multinomial Regression

Coefficients for under-reporting relative to equal-reporting	
Percent Duplicate	2.3
Non-immediate relation to reference person	1.4
Child or grandchild of reference	1.02
Coefficients for over-reporting relative to equal-reporting	
Percent Duplicate	5.9
Proxy reported	2.2
Percent Repeat Visits	2.2

Table 4: GBM Confusion Matrix for Report Status Predictions with Number of Observations Presented

		Reference		
		equal	over	under
Prediction	equal	1145	120	240
	over	23	265	20
	under	182	154	1239

Notes: This table reports how the GBM model categorized the reporting status and compares it to the true value (reference) in the test data set. The diagonal cells report the number of accurate predictions (for example, the first cell reports that the prediction was equal-report status and the reference (true value) was equal-report status). The off-diagonal cells report the number of mis-categorized predictions.

equal- or under-reporting is relatively high but is poor when classifying an individual who is actually an over-reporter. As with the multinomial model, this suggests that different processes are behind whether someone is an over- or under-reporter.

4. Conclusion

While the above analyses are limited by only being able to observe those who reported at least one event during the study period, it provides various insights into the quality of MEPS data, and can inform strategies to improve reporting quality. As seen above, the variables with the largest impact on predicted report status are the paradata variables, and the processes that drive under-reporting are different from the processes that drive over-reporting. The model provides much more accurate predictions for under-reporting, which is more prevalent than over-reporting. Given the large boost to predictive accuracy that comes from using machine learning techniques that are more complex than multinomial regression, utilizing these techniques in survey operations may lead to increased ability to predict which persons are more likely to under or over report their utilization of medical care. Applications of these findings include developing adjustment factors for survey analyses and reallocating resources while the survey is being fielded to maximize data quality. Further research will include analyzing the reporting behavior by event type.

Appendix 1: Figures

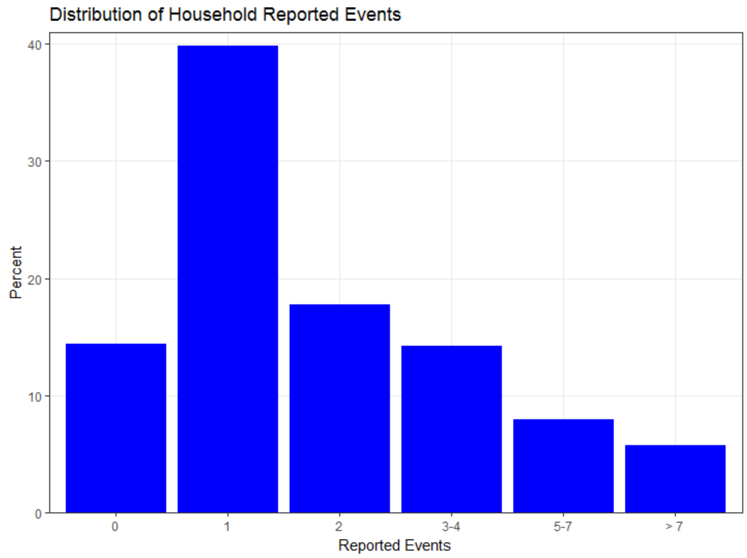

Figure 1: Distribution of household reported events

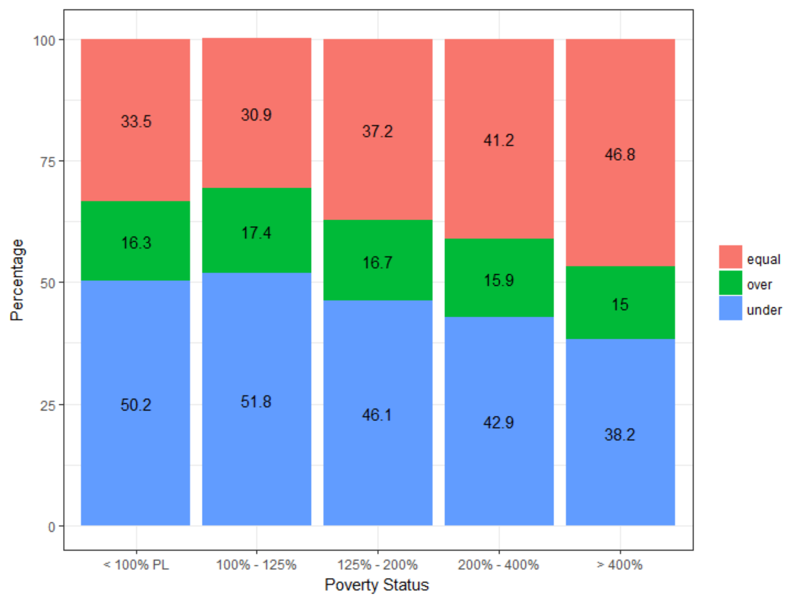

Figure 2: Distribution of reporting status by poverty status

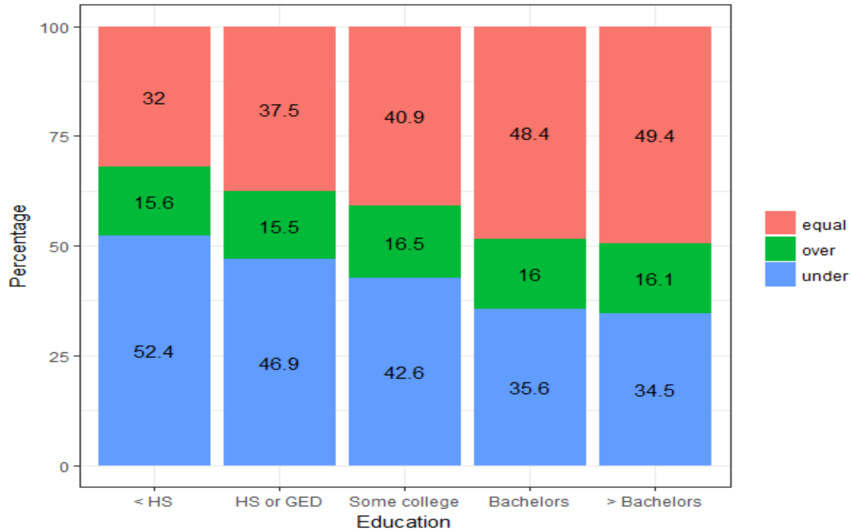

Figure 3: Distribution of reporting status by educational attainment

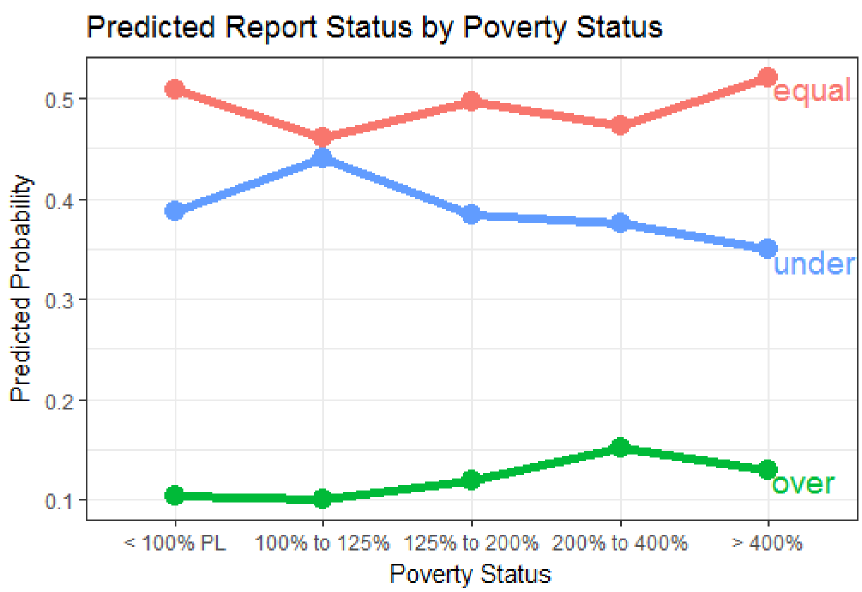

Figure 4: Predicted probability of reporting status by poverty status

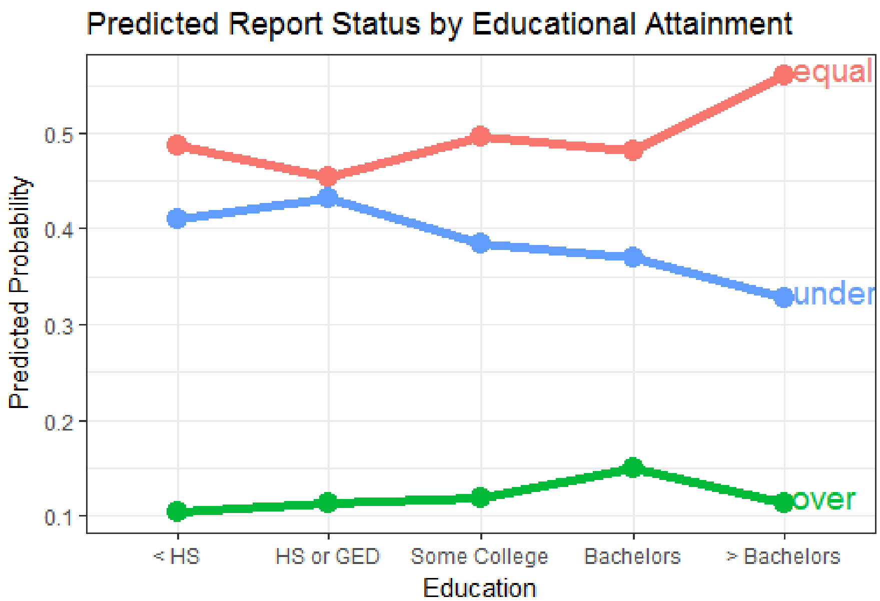

Figure 5: Predicted probability of reporting status by educational attainment

References

Hill, S. & Zuvekas, S. & Zodet, M. (2011). Implications of the Accuracy of MEPS Prescription Drug Data for Health Services Research. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*. 48. 242-59. https://doi.org/10.5034/inquiryjrnl_48.03.04

Machlin, S. & Dougherty, D. (2007). Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey. Rockville, Md.: Agency for Healthcare Research and Quality. https://meps.ahrq.gov/data_files/publications/mr19/mr19.shtml

Zuvekas, S. H., & Olin, G. L. (2009a). Validating Household Reports of Health Care Use in the Medical Expenditure Panel Survey. *Health Services Research*, *44*(5 Pt 1), 1679–1700. http://doi.org/10.1111/j.1475-6773.2009.00995.x

Zuvekas, S. H., & Olin, G. L. (2009b). Accuracy of Medicare Expenditures in the MEPS. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*. *46*(1), 92-108. https://doi.org/10.5034/inquiryjrnl_46.01.92