Fast-Track Estimation Procedures and Analyses to Enhance the Utility of National Survey Data

Steven B. Cohen
RTI International, P.O. Box 12194, Research Triangle Park, NC 27709-2194

Abstract

To meet rigorous statistical quality standards, national surveys often experience a significant lag time from the completion of data collection to the release of the final analytical data files. Consequently, there is a clear demand for the availability of fast-track preliminary/beta versions of the analytical file(s) generated from these ongoing national survey efforts. The availability of preliminary survey estimates and analytic findings based on descriptive and multivariate analyses from these expedited data resources would provide the research and policy community with invaluable insights. These early deliveries would provide signals as to the stability of prior trends or serve as bellwether alerts of likely significant departures or impending issues that could benefit from swift corrective actions.

A substantial amount of time associated with these post data collection statistical tasks is devoted to adjusting for missing data to permit national survey estimates. This is often attributable to the need to anticipate and appropriately adjust for item nonresponse that is nonignorable. When present, imputation strategies that do not account for these conditions will be limited in their capacity to reduce the impact of nonresponse bias on estimates. In this presentation, we examine the performance of fast-track item nonresponse imputation strategies under the assumption of nonignorable item nonresponse in terms of their alignment with final survey estimates on public use files. Attention is given to reverse engineering the imputation process to identify the subset of missing data cases that require specialized treatment. The methodology employed serves to identify patterns of nonignorable item nonresponse and implement corrective strategies to achieve significant efficiencies in terms of cost and time for the production of preliminary analytical files that satisfy well defined levels of accuracy that ensure data integrity. Examples are provided with applications to national survey efforts that include the Medical Expenditure Panel Survey.

Keywords: Efficiency, imputation, modeling

1. Introduction

To meet rigorous statistical quality standards, national surveys often experience a significant lag time from the completion of data collection to the release of the final analytical data files. Consequently, there is a clear demand for the availability of fast-track preliminary/beta versions of the analytical file(s) generated from these ongoing national survey efforts. The availability of preliminary survey estimates and analytic findings based on descriptive and multivariate analyses from these expedited data resources would provide the research and policy community with invaluable insights. These early deliveries would provide signals as to the stability of prior trends or serve as bellwether alerts of likely significant departures or impending issues that could benefit from swift corrective actions.

A substantial amount of time associated with these post data collection statistical tasks is devoted to adjusting for missing data to permit national survey estimates. This is often attributable to the need to anticipate and appropriately adjust for item nonresponse that is nonignorable. When present, imputation strategies that do not account for these conditions will be limited in their capacity to reduce the impact of nonresponse bias on estimates. In this presentation, we examine the performance of fast-track item nonresponse imputation strategies under the assumption of nonignorable item nonresponse in terms of their alignment with final survey estimates on public use files. Attention is given to reverse engineering the imputation process to identify the subset of missing data cases that require specialized treatment. The methodology employed serves to identify patterns of nonignorable item nonresponse and implement corrective strategies to achieve significant efficiencies in terms of cost and time for the production of preliminary analytical files that satisfy well defined levels of accuracy that ensure data integrity. Examples are provided with applications to national survey efforts that include the Medical Expenditure Panel Survey.

2. Project Goal

Demand is increasing for the delivery of fast-track preliminary/beta versions of the analytical file(s) generated from survey data. The survey estimates and preliminary analytic findings based on multivariate analyses conducted by internal research staff that could be derived by these early deliveries would provide analysts with invaluable insights as to the stability of prior trends or serve as bellwether alerts of likely significant departures/impending issues that could benefit from swift corrective actions. For this study, the National Medical Expenditure Panel Survey (MEPS) will be used as the platform for developing the AI solution(s) to generating the fast-track survey estimation and imputed analytic files. The primary objectives of this effort are to achieve reductions in time and cost for client deliverables while achieving data quality standards.

Attention has been given to the imputation process for MEPS to fast track the production of analytical files of acceptable levels of statistical quality and accuracy. For example, the current MEPS imputation process requires substantial time and resources to ensure that data quality thresholds are achieved. This project uses use predictive solutions to determine whether the observed data and imputed data are of acceptable levels of quality to allow the overall process to proceed to analytic file production. These model-based approaches are specified to determine whether quality thresholds are achieved for the resultant survey estimates and, if not, to facilitate adjustments to the imputation process iteratively until acceptable levels of accuracy in estimates are achieved.

3. Development of Fast-Track Analytic Files

The primary focus of this initiative component was the acceleration of the MEPS imputation processes to yield fast-track estimates that serve as early alerts to inform health policy efforts. MEPS is an annual longitudinal national survey that collects data on health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. The survey is sponsored by the Agency for Healthcare Research and Quality (AHRQ). Since its inception, MEPS data have supported a highly visible set of descriptive and behavioral analyses of the U.S. health care system. These include studies of the population's access to, use of, and expenditures and sources of payment for health care; the availability and costs of private health insurance in the

employment-related and non-group markets; the population enrolled in public health insurance coverage and those without health care coverage; and the role of health status in health care use, expenditures, and household decision making, and in health insurance and employment choices. As a consequence of its breadth, the data have informed the nation's economic models and their projections of health care expenditures and utilization. The level of the cost and coverage detail collected in MEPS has enabled public and private sector economic models to develop national and regional estimates of the impact of changes in financing, coverage, and reimbursement policy, as well as estimates of who benefits and who bears the cost of a change in policy.

The evaluation was done in several phases:

- Understand the data.
- Attempt to reproduce the imputation strategy employed in prior cycles of MEPS.
- Modify the off-the-shelf methods.
- Develop promising-in-the-future methods.

To initiate the development of the fast-track imputation estimation methodology for MEPS applications, we concentrated on the medical expenditures and associated sources of payment related to office-based physician visits experienced by the U.S. civilian noninstitutionalized population. The data were further restricted to visits that are not associated with a flat fee or capitation. In examining the current MEPS data, for the 2014 physician-based visits, approximately 50% of the expenditure data are either completely or partially missing.

The first phase of this effort to develop the fast-track imputation strategy required an initial imputation of the missing data using conventional imputation methods, such as weighted sequential hot deck (WSHD). Consequently, analyses were conducted to fit regression models to identify the most salient factors associated with expenditures for physician office visits. These would serve as important imputation class variables. The measures would be prioritized via results from stepwise regression procedures and then recategorized as necessary to define the final imputation class variables. WSHD imputation procedures were then applied to impute the missing payments based on the defined imputation class that is associated with the medical expenses. The quality of the newly imputed data was compared with the complete data and the existing MEPS imputed data via summary statistics and payment distributions.

4. Data Files and Variables

The 2014 MEPS household component (HC) data and office-based medical provider data were downloaded from the AHRQ website at
https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp.
Person-level variables were extracted from the HC; they include demographic, geographic, perceived health status, and insurance coverage variables. Event-level variables were extracted from the MEPS event-level files; they include test procedures performed at the visit, total charge, and various sources of payments. The subset variables from the HC file were merged onto the medical event file by person ID (DUPERSID) to form an initial working dataset for subsequent imputation.

The following payment variables were selected for imputation:

- OBSF14X: AMOUNT PAID, FAMILY (IMPUTED)

- OBMR14X: AMOUNT PAID, MEDICARE (IMPUTED)
- OBMD14X: AMOUNT PAID, MEDICAID (IMPUTED)
- OBPV14X: AMOUNT PAID, PRIVATE INSURANCE (IMPUTED)
- OBVA14X: AMOUNT PAID, VETERANS/CHAMPVA (IMPUTED)
- OBTR14X: AMOUNT PAID, TRICARE (IMPUTED)
- OBOF14X: AMOUNT PAID, OTHER FEDERAL (IMPUTED)
- OBSL14X: AMOUNT PAID, STATE & LOCAL GOV (IMPUTED)
- OBWC14X: AMOUNT PAID, WORKERS COMP (IMPUTED)
- OBOR14X: AMOUNT PAID, OTHER PRIVATE (IMPUTED)
- OBOU14X: AMOUNT PAID, OTHER PUBLIC (IMPUTED)
- OBOT14X: AMOUNT PAID, OTHER INSURANCE (IMPUTED)
- OBXP14X: SUM OF OBSF14X – OBOT14X (IMPUTED)

The charge variable (OBTC14X) on the file was treated as available to define the imputation classes and identify the predictive model. As in MEPS, this variable is imputed prior to the payment variables.

- OBTC14X: HHLD REPORTED TOTAL CHARGE (IMPUTED)

As indicated above, we restricted our data to all respondents with positive weights (PERWT14F>0), visits to physicians only (MPCELIG=1), not a flat fee (FFEEIDX=−1), complete HC and medical provider component (MPC) data, and fully or partially imputed data (IMPFLAG [1] =1,2,3,4). Only fully imputed medical expenditures (where IMPFLAG=3) were considered for re-imputation in this analysis.

5. Assessing Convergence in Expenditure Distributions among Population

The following diagnostic criteria were used to assess the quality and accuracy of the imputation procedures:

statistical tests to assess the convergence in the use and expenditure estimates between the fast-track and existing MEPS imputed estimates,

statistical tests to assess the convergence in the estimated medical expenditure distributions and their concentration between the fast-track and existing MEPS imputed estimates, and

assessments of the alignment of statistically significant measures in analytic models predicting medical expenditures

The first Figure illustrates the alignment in the source of payment estimates derived from our fast track method in comparison with the final estimates on the MEPS public use files. We also assessed convergence in the estimated medical expenditure distributions and their concentration between the fast-track and existing MEPS imputed estimates. Table 5-1 demonstrates the convergence in distributional estimates of person-level medical expenditures based on the fast-track imputation strategy for 2014. Specific to the overall payment variable, this was implemented by calculating the distribution of total payments among the population. First, the event payment data, restricted to not-a-flat-fee visits to physician providers only, were aggregated to the person-level data. Then, using the weights, we determined the percentage of overall office-based expenditures consumed by the top

[1] Imputation status in the MEPS office-based medical provider visits data, 1 = complete HC data, 2 = complete MPC data, 3 = fully imputed data, and 4 = partially imputed data. Values 0 (not eligible for imputation) and 5 (capitation imputation) are not considered in this analysis.

1%, 5%, 10%, 20%, 25%, 30%, 40%, and 50% of the population with office-based visits. In addition, the mean expenses for each of these percentiles and their SEs were calculated.

Means and Standard Errors of the Medical Expenditures of Visits to Physicians by Existing Data and Weighted Sequential Hot-Deck Imputed Data with Adjustment Made to the Top 5% of Most Divergent Records, 2013–2014 MEPS

	2014 Existing Data (n=120,893)			2014 First Pass WSHD Imputed Data (n=120,893)			2014 First Pass WSHD Imputed Data with Z(i)-Based Adjustment (n=120,893)			2013 Existing Data (3) (n=120,189)		
	Unweighted	Weighted		Unweighted	Weighted		Unweighted	Weighted		Unweighted	Weighted	
Expenditure	Mean	Mean	SE Mean	Mean	Mean	SE Mean	Mean	Mean	SE Mean	Mean	Mean	SE Mean
Amount Paid By												
Family	21.19	25.98	0.95	28.43	31.85	0.81	27.26	30.81	0.77	20.90	24.79	0.74
Medicare	54.72	57.30	3.04	55.07	57.08	2.87	53.61	55.07	2.72	48.07	52.16	2.41
Medicaid	30.32	18.88	1.08	27.24	17.39	0.97	27.18	17.79	1.06	35.01	20.83	1.40
Private Insurance	75.59	87.02	3.35	78.66	88.28	3.05	76.68	86.20	3.01	80.29	95.34	4.42
Veterans/CHAMPVA	6.71	6.48	1.13	1.32	1.30	0.54	2.59	2.64	0.68	5.10	4.72	0.56
TRICARE	1.94	1.99	0.44	1.77	1.81	0.38	1.86	2.01	0.40	2.22	2.81	0.60
Other Federal	0.64	0.52	0.20	0.14	0.06	0.03	0.21	0.19	0.08	0.58	0.48	0.11
State & Local Gov	3.04	1.82	0.34	1.54	1.05	0.17	1.88	1.39	0.20	6.03	4.02	0.83
Workers Comp	3.37	2.37	0.34	1.22	1.19	0.24	1.88	1.47	0.24	3.87	3.26	0.67
Other Private	4.58	5.21	1.09	5.02	5.08	0.92	5.03	5.01	0.91	4.65	4.05	0.51
Other Public	0.56	0.30	0.05	0.50	0.33	0.07	0.49	0.34	0.07	0.93	0.63	0.12
Other Insurance	3.67	3.00	0.43	2.93	2.50	0.32	3.34	2.90	0.37	5.38	4.64	1.06
Total Paid	206.33	210.86	3.85	203.84	207.92	3.63	201.99	205.82	3.61	213.03	217.72	5.07

Notes: MEPS = Medical Expenditure Panel Survey; WSHD = weighted sequential hot-desk; SE = standard error

This table was produced based on data that passed through the first pass imputation and Z(i)-based adjustment with a whole vector of sources of payment replaced in the 2014 data imputation.

Table 5-1: Person-Level Comparison of Percentage of the Total Expenditures and Mean Expenditures among the Population Between Actual Office-Based Physician Visit Event Data and Fat Track Hot-Deck Imputed Data (n=21,399), 2014 MEPS

Top Percentile, %	Actual Data				Model-Informed Imputation			
	Percent	SE Percent	Mean	SE Mean	Percent	SE Percent	Mean	SE Mean
1	21.66	1.42	27,906	1,234	20.47	1.45	25,691	1,066
5	43.92	1.27	11,327	383	42.41	1.30	10,682	354
10	57.46	1.07	7,413	213	56.33	1.07	7,093	198
20	72.95	0.74	4,704	115	72.21	0.76	4,547	110
25	78.14	0.62	4,033	96	77.52	0.64	3,905	90
30	82.26	0.50	3,538	83	81.73	0.52	3,431	78
40	88.33	0.37	2,849	63	87.96	0.37	2,769	61
50	92.51	0.24	2,387	54	92.25	0.24	2,323	52

Note: MEPS = Medical Expenditure Panel Survey; SE = standard error

We also assessed the alignment of statistically significant measures in analytic models predicting medical expenditures. Table 5-2 presents results of the fast-track imputation procedure applied to 2014 MEPS data comparing the correspondence of significant predictors of individuals experiencing the highest 5% aggregated totals of medical expenditures for office-based physician visits. As shown, the beta coefficients, their SEs and their level of significance derived from the fast-track imputations are aligned with those derived from the actual imputed data.

Table 5-2. : Logistic Regression Comparison for Individuals Likely to Be on the Top 5% of the Total Health Care Expenditure Distribution Using the MEPS Data Restricted to Office-Based Physician Provider Visits and Hot-Deck Imputed Data with Adjustment Made to the Top 5% of Most Divergent Records (n=21,399), 2014 MEPS

	MEPS Actual Data (R^2=0.1201)			WSHD Imputed Data (R^2=0.1211)		
Measures	Beta Coefficient	SE of Beta	Wald F *P*-Value	Beta Coefficient	SE of Beta	Wald F *P*-Value
Age	−0.0013	0.0045	0.7719	−0.0017	0.0045	0.7089
Sex						
Male	0.0000	0.0000	0.0126	0.0000	0.0000	0.0161
Female	−0.2777	0.1103		−0.2667	0.1099	
Race/Ethnicity						
Hispanic	0.0000	0.0000	0.7521	0.0000	0.0000	0.9380
Non-Hispanic White	−0.0424	0.1424		−0.0049	0.1387	
Non-Hispanic Black	0.1070	0.1685		0.0302	0.1646	
Non-Hispanic Other	0.0335	0.2231		−0.1034	0.2391	
Marital Status						
Married	0.7912	0.2640	0.0045	0.6742	0.2638	0.0142
Widowed	1.2668	0.3508		1.2483	0.3614	
Divorced/Separate	1.1008	0.3132		1.0272	0.3403	
Never Married	0.8880	0.2720		0.6707	0.2787	
Under 16	0.0000	0.0000		0.0000	0.0000	
Family Size						
One	0.0000	0.0000	0.1440	0.0000	0.0000	0.4124
Two or more	0.2414	0.1646		0.1435	0.1747	
Region						
Northeast	0.0000	0.0000	<0.0001	0.0000	0.0000	<0.0001
Midwest	0.5280	0.1568		0.5364	0.1675	

South	−0.3967	0.1312		−0.3164	0.1477	
West	0.0020	0.1466		0.1650	0.1419	
Family Income Classification						
Poor	0.0000	0.0000	0.3635	0.0000	0.0000	0.3108
Near Poor	0.1864	0.3322		0.1168	0.3495	
Low Income	−0.0195	0.2046		−0.0566	0.2121	
Middle Income	0.1956	0.1936		0.1395	0.1864	
High Income	0.3088	0.2076		0.3398	0.2115	
Health Insurance Coverage						
Any private	0.3705	0.3416	0.5182	0.1098	0.2986	0.3241
Public only	0.2610	0.3220		−0.1298	0.2793	
Uninsured	0.0000	0.0000		0.0000	0.0000	

(continued)

	MEPS Actual Data (R^2=0.1201)			WSHD Imputed Data (R^2=0.1211)		
Measures	Beta Coefficient	SE of Beta	Wald F *P*-Value	Beta Coefficient	SE of Beta	Wald F *P*-Value
Health Status						
Excellent	0.0000	0.0000	0.0085	0.0000	0.0000	0.0669
Very Good	0.4055	0.1818		0.2231	0.1756	
Good	0.6591	0.1821		0.4645	0.1776	
Fair	0.5461	0.2336		0.5361	0.2252	
Poor	0.6753	0.3085		0.2686	0.3171	
Limitation in Activity						
Yes	0.0763	0.1907	0.6896	0.2017	0.1873	0.2828
No	0.0000	0.0000		0.0000	0.0000	
Cancer						
Yes	0.5212	0.1610	0.0014	0.4011	0.1703	0.0195
No	0.0000	0.0000		0.0000	0.0000	
Heart Disease[a]						
Yes	−0.1347	0.1452	0.3546	−0.2135	0.1563	0.1735
No	0.0000	0.0000		0.0000	0.0000	
High Blood Pressure						
Yes	−0.2524	0.1312	0.0558	−0.2586	0.1445	0.0751
No	0.0000	0.0000		0.0000	0.0000	
Inpatient Events	0.0676	0.0725	0.3524	0.0558	0.0761	0.4644
Number of Prescribed Medicine Purchases	0.0029	0.0022	0.1769	0.0035	0.0022	0.1050
Number of Ambulatory Visits	0.1770	0.0075	<0.0001	0.1814	0.0077	<0.0001

Notes: MEPS = Medical Expenditure Panel Survey; n = sample size; SE = standard error.

The 2014 MEPS Household Component data and Medical Provider Component data were downloaded from the following website: https://meps.ahrq.gov/data_stats/download_data_files.jsp. This analysis was restricted to physician office visits only (MPCELIG=1) where there was not a flat fee (FFEEIDX=−1) and all respondents have positive weights with completed or partially/fully imputed HC and MPC data (IMPFLAG=1,2,3,4).

[a] The heart disease was defined as "Yes" if a respondent was diagnosed with coronary heart disease, heart attack, or other heart disease.

6. Summary

This effort has focused on identifying and implementing fast track estimation and imputation procedures. The objective was to fast track the generation of survey estimates from national surveys prior to data collection completion and final analytic data file production while satisfying well-defined levels of accuracy and ensuring data integrity. This capability would (1) satisfy demand from current and future clients for early alerts regarding new trends and unexpected findings; (2) automate manual tasks by using input data and establishing predefined outcome preferences; (3) permit the user to focus energy on higher-order problem resolution; and (4) achieve gains in timeliness, cost, and quality in final survey products by the earlier identification and resolution of estimation and imputation issues that have surfaced.

The fast track applications to the MEPS imputation process uncovered underlying structures to the final data on the public use files produced. The final results were achieved by a hybrid approach that combined weighted sequential hot-deck methods with predictive modeling. In this study, the fast track methods we employed yielded comparable survey estimates relative to those produced from the MEPS final imputed data, which we considered the "gold standard."

Acknowledgements

Special acknowledgements go to Feng Yu, Jamie Shorey and Georgiy Bobashev, RTI International for their contributions.

References

Abayomi, K., Gelman, A., & Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society Series C-Applied Statistics, 57*(Part 3), 273–291. doi:10.1111/j.1467-9876.2007.00613.x

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association, 112*(518), 859–877. doi:10.1080/01621459.2017.1285773

Cohen, S. B. & J. Cohen, 2013. "The Capacity of the Medical Expenditure Panel Survey to Inform the Affordable Care Act", Inquiry. 50(2):124-34

Cohen, J., S. Cohen, and J. Banthin. 2009. "The Medical Expenditure Panel Survey: A National Information Resource to Support Healthcare Cost Research and Inform Policy and Practice." Medical Care 47 (7, Suppl. 1): 44–50.

Cohen, S., and T. Buchmueller. 2006. “Trends in Medical Care Costs, Coverage, Use and Access: Research Findings from the Medical Expenditure Panel Survey.” Medical Care 44 (5): 1–3.

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. *Proceedings of the American Statistical Association*, 721–726.De Jongh, M., & Druzdzel, M. F. (2009). A comparison of structural distance measures for causal Bayesian network models. *Recent Advances in Intelligent Information Systems*. 443–456.

Goodfellow, I. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*.

Gregor, K. (2015). DRAW: A recurrent neural network for image generation. *Advances in Neural Information Processing Systems*.

Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics, 20*(8), 897–916. doi:10.1002/hec.1653

Mohan, K., Pearl, J., & Tian, J. (2013). Missing data as a causal inference problem. Technical Report R-410: UCLA.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro & W. J. Frawley (Eds.), *Knowledge Discovery in Databases*. Cambridge, MA.: AAAI/MIT Press.