Artificial Intelligence (AI)-Enhanced Applications to Survey-Specific Imputation Tasks to Achieve Time and Cost Efficiencies

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

Abstract
A high degree of rigor is essential in the statistical integrity of "end-product" analytic resources that are used to inform policy and action. In this vein, statistical and analytic staff devote substantial time and effort to implement estimation and imputation tasks; these tasks are essential components of the "end-product" analytic databases derived from national or sub-national surveys and related data collections. These efforts require a substantial commitment of project related funds to achieve, and significant lag times often exist from the time data collection is completed to the time the final analytical data file is released. This paper focuses imputation methodology enhanced with artificial intelligence (AI) for specific national survey efforts. We demonstrate the efficiencies achieved by the AI-enhanced applications in terms of cost and time that satisfy well-defined levels of accuracy to ensure data integrity. Attention is given to AI-enhanced processes that serve as an alternative solution to manual, repetitive or time-intensive tasks. Examples are provided with applications to national survey efforts that include the Medical Expenditure Panel Survey.

Keywords: imputation; artificial intelligence; survey efficiencies; MEPS

1. Introduction

A high degree of rigor is essential in the statistical integrity of "end-product" analytic resources that are used to inform policy and action. In this vein, statistical and analytic staff devote substantial time and effort to implement estimation and associated imputation tasks, which are essential components of the "end-product" analytic databases derived from national or sub-national surveys and related data collections. These efforts require a substantial commitment of project related funds to achieve and significant lag times often exist from the time data collection is completed to the time the final analytical data file is released. This paper focuses on the development and implementation of artificial intelligence (AI) and machine learning enhanced applications to imputation for specific national survey efforts that achieve efficiencies in terms of cost and time while satisfying well defined levels of accuracy that ensure data integrity. Attention is given to enhanced processes that: serve as an alternative solution to manual, repetitive or time-intensive tasks; operationalize decisions based upon predefined outcome preferences and upon access to input data that sufficiently informs the decisions; facilitate real-time interpretation and interactions for accessing and acting upon the AI-derived decisions to permit the user to focus energy on higher-order thinking and problem resolution. Our approach includes the framing of predictions of criterion variables and their distributions as a multi-task learning (MTL) problem. MTL jointly solves multiple learning tasks by exploiting the correlation structure across tasks. Consideration is also given to the application of random forest methods which utilize an ensemble of decision trees to facilitate predictions.

Examples are provided with applications to national survey efforts that include the Medical Expenditure Panel Survey (MEPS). MEPS is a large scale annual longitudinal national based survey that collects data on health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. This research effort focuses on harnessing AI/ML techniques to yield MEPS expenditure data and estimates that are closely aligned with the actual results that require several months to produce and are provided in the MEPS final analytic files. The methods performance is evaluated based on the medical expenditure data sets released as public use files, which are regarded as the reference standard in the evaluation phase of this study.

2. Project Goal

Statistical and analytic staff devote substantial time and effort to implementing the estimation and imputation tasks that are essential components of the end-product analytic databases derived from national or sub-national surveys and related data collections. Clients demand a high degree of rigor in the statistical integrity of these end-product analytic resources that are used to inform policy and action. To achieve the targeted level of quality in the final estimation weights and imputation procedures for critical key analytic measures and other core survey data elements, a very significant time lag occurs from completion of data collection to release of the final analytic data file.

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 may 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 AI and machine learning (ML) derived 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 AI/ML derived 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.

MEPS consists of a family of three interrelated surveys: Household Component (MEPS-HC), Medical Provider Component (MEPS-MPC), and Insurance Component (MEPS-IC). MEPS-IC also collects establishment-level data on insurance programs. Through a series of interviews with household respondents, MEPS-HC collects detailed information at the level of the individual respondent on demographic characteristics, health status, health insurance, employment, and medical care use and expenditures. These data support estimates both for individuals and for families in the United States. Respondents identify medical providers from whom they have received services [3-5]. The set of households selected for MEPS-HC is a subsample of 15,000 households/35,000 individuals participating in the National Health Interview Survey (NHIS). The MEPS-HC survey consists of an overlapping panel design in which any given sample panel is interviewed a total of five times in person over 30 months to yield annual use and expenditure data for 2 calendar years. These rounds of interviewing are conducted at about 5- to 6-month intervals. They are administered through a computer-assisted personal interview mode of data collection and take place with a family respondent who reports for him/herself and for other family members. Data from two panels are combined to produce estimates for each calendar year. Westat is the data collection organization for MEPS-HC.

MEPS-MPC is a survey of the medical providers, facilities, and pharmacies that provided care or services to sample persons. The primary objective is to collect detailed data on the expenditures and sources of payment for the medical services provided to individuals sampled for MEPS. Such data are essential to improve the accuracy of the national medical expenditure estimates derived from MEPS, given that household respondents are not always the most reliable sources of information on medical expenditures. MPC data are collected a year after the household health care event information is collected to allow adequate time for billing transactions to be completed. MPC collects data on dates of visits/services, use of medical care services, charges, sources of payments and amounts, and diagnoses and procedure codes for medical visits/encounters. Only providers for whom a signed permission form was obtained from the household authorizing contact are eligible for data collection in MPC. The categories of providers in MPC include (1) office-based medical doctors; (2) hospital facilities providing inpatient, outpatient, and emergency room care; (3) health maintenance organizations (HMOs); (4) physicians providing care during a hospitalization; (5) home care agencies; and (6) pharmacies. RTI International is the data collection organization for MEPS-MPC.

This effort focused on employing AI/ML techniques to yield imputed MEPS expenditure data that are aligned with the results that required several months to produce in order to release the MEPS final analytic Public Use files. The method's performance was evaluated based on the AHRQ-derived imputed dataset, which was regarded as the reference standard.

The evaluation was done in several phases:

- Understand the data.
- Attempt to reproduce the imputation strategy employed in prior cycles of MEPS.
- Evaluate the off-the-shelf AI/ML methods.
- 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 at the Population Level

We 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%, 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. These tables indicate that the estimated medical expenditure distributions and their concentration between the fast-track and existing MEPS imputed estimates have a good level of alignment. Note that the number of observations used for these estimates is generally less than the number of person-level records with positive weights from the HC file because they are only a subset of the event data that were restricted to not-a-flat-fee office-based physician visits in the fast-track imputation.

[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.

Table 5-1: Person-Level Comparison of Percentage of the 2014 Total Expenditures and Mean Expenditures among the Population: Public Use File Office-Based Physician Visit Event Data and Fast Track Hot-Deck Imputed Data aggregated to the person level.(n=21,399), 2014 MEPS

	Public Use File Data				**Fast Track Informed Hot-Deck Imputation**			
Top Percentile, %	**Percent of total $**	**SE Percent**	**Mean**	**SE Mean**	**Percent of total $s**	**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

6. An AI Approach to Fast Track MEPS Imputation

Armed with a better understanding of the nuances and impact of alternative fast-track MEPS imputation strategies, we then demonstrated the use of ML algorithms for estimating health care expenditures for application to MEPS. Our predictor variables included basic demographic information, categorized insurance costs for current year, and more than 80 condition and provider categories, listing medical conditions and provider variables. To tackle the large number of covariates and the highly nonlinear nature of health care costs, we used hierarchical statistical regression methods. We also investigated the use of Classification and Regression Trees (CART) and Random Forests (RF) to estimate unknown variables related to health care costs for office-based provider visits. We demonstrated that (a) ML approaches can approximate the standard imputation process in much shorter time; (b) although ML algorithms are also limited by skewed cost distributions in health care, for a large fraction of health care events within the population, we were able to predict with higher accuracy using these algorithms; and (c) our methods can also be used to evaluate future costs for segments of the population with reasonably low error. Our analysis shows that RF is a promising method for predictive modeling, providing the best performance across a range of other regression methods we tried.

6.1 Multi-Output Random Forest

The prediction of medical expenditure composition can be framed as a multi-task learning (MTL) problem. MTL jointly solves multiple learning tasks by exploiting the correlation structure across tasks. Let us denote $y = ([y]_k) \in \mathbb{R}^K$ as the K target variables to be predicted and $x \in \mathbb{R}^p$ as the predictors. The joint probability of variables writes $p(y, x)$. In the statistical learning framework, the marginal likelihood $p(y) = \int p(y, x)\mathrm{d}x = \int p(y|x)p(x)\mathrm{d}x$ is maximized, where the integration over $p(x)$ is replaced with the summation over empirical distribution in actual practice. The correlation structure of y essentially comes from two sources: (1) the shared x and (2) the conditional correlation $y|x$. Intuitively, incorporating other tasks provides additional supervision to the learner, which translates into better feature extraction and sample loss during the training phase. This holds true unless the predictive features for $[y]_k$ do not overlap and $[y]_k$s are statistically independent given x. In the MEPS study, the amount of payment made by different sources is correlated, and therefore using MTL is a natural choice.

We choose *Multi-Output RF* as our regressor. RF builds on aggregating the prediction from an ensemble of decision trees, and it has a proven record in survey data analysis. Decision trees are independently trained with a bootstrap sample of the training data, often referred to as the *bagging* technique. Each tree is constructed in a recursive fashion. A subsample of the training data falls on a tree node, and a (possibly random) subset of the features is selected. Then, for each feature, the algorithm enumerates all possible splitting points (decision boundaries) and computes an impurity score based on the splitting. The impurity scores, usually the information gain or Gini score, reflect the homogeneity of the sample given the split. If the splitting condition is satisfied, the best split based on the impurity score will be executed, which makes the node a decision node; otherwise, the node becomes a leaf node. For multi-output, the impurity score is usually computed on a task-based fashion and then aggregated to make the decision. The final ensemble averaging step turns a swarm of diversified, potentially unstable, weak learners into a robustness strong learner. The impurity score, on which the splitting rule hinges, usually does not rely on a particular statistical model. This adds to the robustness of decision tree [1-2, 6-11].

6.2 Results

The ML modeling approach was iteratively developed and tested on several years of prior MEPS event level source of payment data, using the fully known rows and then the values for the source of payment vector were estimated using a trained RF model. Results for within-year predictions on a more recent MEPS event level expenditure dataset not used in the model development process are presented in Table 6-1. Both weighted and unweighted means are shown and demonstrate good alignment with the expenditure data on the public use files. The largest errors were in the Private source of payment category, followed by Workers Compensation as a source of payment.

Table 6-1: Comparisons of RF Fast-track Derived and Public Use File Mean Office Based Physician Event Expenditure Estimates by Source of Payment, 2015 MEPS

	Mean		**Weighted Mean**		
Source of Payment	**PUF File Imputed $s**	**RF Predicted $s**	**PUF File Imputed $s**	**RF Predicted $s**	**Difference ($)**
Medicaid	31.5	31.84	20.26	20.66	0.4
Medicare	58.18	58.04	60.34	60.2	-0.15
Other Federal	0.58	0.65	0.66	0.73	0.07
Other Private	4.97	4.99	4.33	4.38	0.05
Other Insurance	4.28	4.25	4.62	4.6	-0.02

Other Public	0.76	0.74	0.51	0.51	0
Private Insurance	83.33	82.84	98.11	97.54	-0.58
Family	22.97	22.98	27.87	27.87	0
State & Local Gov	1.18	1.2	1.13	1.12	-0.01
TRICARE	2.03	2.09	2.35	2.4	0.05
Workers Compensation	4.07	4.24	4.24	4.41	0.17
Veterans/CHAMPVA	8.04	8.12	8.52	8.61	0.09
Sum of OBSF12X–OBOT12X	221.9	221.99	232.90	233.03	0.06

7. Summary

This effort has focused on identifying and implementing AI-based applications to 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 AI 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 statistical profile matching and high-level AI (RF)-based imputation. For several of the years under study, the AI-based methods we employed yielded comparable survey estimates relative to those produced from the MEPS final imputed data, which we considered the "gold standard." Future applications of this methodology have the capacity to yield significant reductions in the time and cost in the production of preliminary analytic files that could help provide policymakers with early alerts of significant departures/impending issues that could benefit from swift corrective actions.

Acknowledgements

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

References

[1] 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

[2] 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

[3] 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

[4] 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.

[5] 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.

[6] 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.

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

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

[9] 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

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

[11] 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.