Imputation of Medicaid Beneficiaries' Expenditures for Physician Visits and Hospital Care in the Medical Expenditure Panel Survey

Jeffrey M. Gonzalez* Lisa B. Mirel

Abstract
The Medical Expenditure Panel Survey (MEPS) collects data on health care expenditures for the US civilian noninstitutionalized population. MEPS includes two key components: the household (HC) and the medical provider component (MPC). The HC was designed to enumerate all health care service events over a two-year period and their associated charges and payments for all sample persons via a series of five personal-visit interviews. However, the household is not always the best source of information on medical expenditures. This is particularly true for enrollees in the Medicaid program, where financial transactions occur between the medical provider/plan and the state Medicaid agency. The MPC was designed to supplement household reported data with health care provider data for a subset of the household reported events (e.g., physician visits and hospital care). In this paper, we describe the methodology used to impute missing expenditure information for Medicaid beneficiaries in MEPS. We then examine the extent to which Medicaid and out-of-pocket payments for reported health care service events of Medicaid beneficiaries are obtained via the household, medical providers, or imputation. Finally, we identify areas of future research.

Key Words: Data quality, Measurement error, Total survey error, Underreporting, Imputation

1. Introduction and motivation

The Medical Expenditure Panel Survey (MEPS) is the most complete data source on the cost and use of health care and insurance coverage in the United States (US). It is a nationally representative survey of US households and is used by various stakeholders, such as academic researchers and policymakers, to describe and analyze the levels and determinants of health care utilization and spending by US households and to inform health care policy. Effective use of these data for statistical purposes and policy decisions requires that the data are of high quality, i.e., they are relevant, accurate, timely, complete, etc.

However, a pervasive problem in household surveys requiring respondents to perform recall tasks and report on events and characteristics associated with those events is that errors in the retrieval process can affect the quality of the data collected (Groves et al., 2004). MEPS is a prime example of this type of survey because it requires household informants to report all health care service events and their associated charges and payments for themselves and their family members over a specific time period. Because of the potential difficulties with this recall task, MEPS respondents are encouraged to maintain records (e.g., Explanation of Benefits reports from private health care insurance companies) and consult them when reporting their family's health care utilization and costs (Machlin et al., 2010). Enrollees in the Medicaid program, however, do not have the luxury of consulting these types of records since financial transactions for the services received typically occur between the provider/plan and the state Medicaid agency. Because of these arrangements, the state Medicaid agency may never send the Medicaid beneficiary an Explanation of Benefits report. This may inhibit the Medicaid beneficiary's (or family member) ability to perform adequately the required recall and reporting tasks.

The Medicaid program is a social health care program managed by individual states, but jointly funded by the individual states and the federal government (Centers for Medicare

*jeffrey.gonzalez@ahrq.hhs.gov, Agency for Healthcare Research and Quality, Rockville, MD

and Medicaid Services, 2014). It provides health care coverage to vulnerable populations in the US, such as families and individuals with low income. Eligibility for Medicaid varies by state, but The Affordable Care Act of 2010 created a national Medicaid minimum eligibility level of 138% of the federal poverty level for nearly all Americans under age 65 (Centers for Medicare and Medicaid Services, 2014). However, in June 2012, the US Supreme Court ruled that states may opt out of expanding Medicaid, and as of June 2014, 21 states have done so (including three that are still considering expanding Medicaid) (The Henry J. Kaiser Family Foundation, 2014). Additionally there are other population subgroups for which persons meeting criteria for inclusion in these subgroups are eligible to receive Medicaid benefits, provided that they pass an income and/or asset test. Examples of these subgroups include non-elderly individuals with disabilities who are employed or seeking employment and pregnant women. In 2011, 20% of the US civilian, noninstitutionalized population age 65 and under were covered by only public sources of health insurance coverage (e.g., Medicaid) (Agency for Healthcare Research and Quality, 2013d).

Since the Medicaid program is managed by individual states and beneficiaries may not receive records for their health care services from the Medicaid program, Medicaid beneficiaries are a targeted domain for MPC data collection in order to supplement their household reported data with medical provider data. Specifically, when a respondent reports having Medicaid coverage during the interview, the MEPS protocol is such that payment and charge information for household reported physician visits tends to have a higher probability of being collected via medical providers. Furthermore, even if a respondent reports non-zero dollar Medicaid payments for physician visits and hospital care, due to the lack of confidence in these reports, MEPS processing procedures are such that the reported Medicaid payment information is generally ignored and subsequently replaced with either medical provider data or through imputation.

In this paper, we examine the extent to which payment information for Medicaid beneficiaries' physician visits and hospital care on the MEPS Public Use Files is obtained from the household, medical provider, or through imputation. To accomplish this, we first describe the imputation procedures utilized in the MEPS program to derive health care expenditure estimates for Medicaid beneficiaries. We then compare the distribution of data source (HC, MPC, or imputation) for Medicaid beneficiaries' Medicaid payments to that of out-of-pocket (OOP) payments for the same event types. Finally, we discuss some implications of these procedures for survey design and estimation and identify potential areas of future research.

2. MEPS background

2.1 MEPS description

MEPS collects data on health care utilization, expenditures, sources of payments, and health insurance coverage for the US civilian noninstitutionalized population. MEPS includes two key components: the household (HC) and the medical provider component (MPC). The HC was designed to enumerate all health care service events over a two-year period and their associated charges and payments for all sample persons while the MPC was designed to supplement household reported data with health care provider data for a subset of the household reported events (e.g., office-based physician visits, hospital care, and home healthcare agency events).

2.1.1 Household component (HC)

The HC is a complex, multi-stage, nationally representative sample of the US civilian, non-institutionalized population. It has been an annual survey since 1996. Each year a new sample is drawn as a subsample of responding households to the previous years' National Health Interview Survey (NHIS), conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. In 2011, the respondent sample size was 33,662 (from 13,449 households) with a response rate of 54.9%. This response rate includes NHIS nonresponse and nonresponse for each round of data collection in MEPS (Agency for Healthcare Research and Quality, 2013a).

The HC employs an overlapping panel design and data are collected using a computer assisted personal interviewing instrument on five separate occasions covering a cumulative two-year reference period. Data are typically collected via one respondent per household, so for multi-person households, the respondent is not always reported for only him/herself. The types of health care events, subsequently referred to as *event type*, inquired about in the HC include inpatient hospital stays, emergency room visits, outpatient department visits, office-based physician visits, dental care, short-term institutional care (e.g., psychiatric care), prescription medications, and other medical supplies (e.g., eyeglasses and insulin). In this paper, we focus on hospital care (e.g., inpatient hospital stays, emergency room visits, and outpatient department visits) and office-based physician visits.

Whenever possible, data on charges and payments for all reported health care service events are collected from the household. Expenditure data on payments are reported as being paid by one of the following "source of payment" categories: (1) family/self OOP; (2) Medicare; (3) Medicaid; (4) Private insurance; (5) Veterans Administration; (6) TRICARE (assistance to dependents of military personnel); (7) Other federal programs (e.g., Indian Health Service, Military treatment facilities, federally-funded programs other than Medicaid); (8) Other state and local programs (e.g., community neighborhood clinics, state and local health departments, state programs other than Medicaid); (9) Workers compensation; and, (10) Other sources (e.g., automobile, homeowners or liability insurance payments).

There are two additional points worth noting about the design of the HC. First, a post-data collection processing step is executed so that data on "total payments (from all sources)" for the health care service event are computed by summing the payment amounts from the individual source of payment categories. Second, health insurance coverage is based on respondent reports. The respondent reports for specific sources of coverage and these reports are then used in the imputation and estimation procedures.

2.1.2 Medical provider component (MPC)

The MPC collects data on dates of visits and services, use of medical care services, charges and sources of payments and amounts, and diagnoses and procedure codes for medical visits from a sample of medical care providers (e.g., physicians, hospitals, home health agencies, and pharmacies) who provided medical care to HC sample members during their visits. We note that eligibility related to ambulatory visits for the MPC is restricted to services rendered in a hospital or by a medical doctor (MD) or doctor of osteophathy (OD) or under the supervision of an MD or OD. We refer to these types of health care services as *MPC-eligible events*. Thus, services rendered at dental offices, for example, are not eligible for inclusion into the MPC.

It is important to note that the MPC is not designed as a nationally representative sample of providers and due to budgetary constraints it only includes a subsample of office-based physicians while all hospitals that are reported as the site of care for inpatient hospital stays, outpatient department visits, and emergency room encounters for HC sample persons are included in the MPC. Furthermore, medical providers only have the potential of being contacted if a HC respondent identifies them as providing care for a reported event and if the respondent signs an authorization form giving permission to MEPS staff to contact and collect the information from the medical providers. Even if the respondent signs an authorization form, a medical provider can still refuse to participate in the MPC. In 2011, the final eligible sample of providers was 45,096 coming from 66,910 eligible sample pairs[1] wtih pair-level response rates varying by provider type (e.g., 90.9% for hospitals and 88.7% for office-based physicians) (Agency for Healthcare Research and Quality, 2013b; 2013c).

2.1.3 Linkage of the two components

Once data are collected from both the HC and MPC, a prerequisite step to the MEPS expenditure estimation methodology is to match medical provider reported information to the appropriate household reported health care events. The linkage between the HC and MPC uses a probabilistic matching procedure based on the Fellegi and Sunter algorithm (Fellegi and Sunter, 1969). Specific details of this matching procedure are ancillary to this report, but can be found in Mirel and Machlin (2013).

3. Imputation procedures

When matching of household and medical provider reports is complete, the general strategy for deriving MEPS expenditure estimates (for MPC-eligible events) is to:

1. Use medical provider reported information whenever possible;
2. In the absence of MPC data, use household reported data from the HC, if complete, and with any necessary adjustments and edits for inaccuracies; and,
3. In the absence of both HC and MPC data, impute missing expenditure information.

In 2010, MEPS switched from weighted sequential hot-deck (See Cox 1980 for a general reference on weighted sequential hot-deck) to predictive mean matching (See Little 1988 for a general reference on predictive mean matching) to impute missing payment information for MPC-eligible events. In 2011, the use of predictive mean matching was expanded to the imputation of missing payment information for other event types. Following, we provide a general overview of predictive mean matching, discuss MEPS' implementation of predictive mean matching, and identify specific procedures related to the imputation of health care expenditures for Medicaid beneficiaries.

3.1 General overview of predictive mean matching

Predictive mean matching is a "nearest neighbor" imputation technique that involves defining a distance function on a covariate space, evaluating it at the values for the recipients (i.e., records with missing values or item nonresponse) and donors (i.e., records with complete information), and then transferring the values from the "nearest" or "closest" donor to the recipient. These transferred values are the imputed values. The choice of distance function typically depends on the nature of the imputation problem, but when it is based on a parametric model, such as a regression model, then this technique is often referred to as *predictive mean matching* (Little, 1988).

[1]The sample unit for the MPC is actually the combination of the provider and the patient (i.e., provider-patient pair) as one provider may have multiple patients.

In mathematical terms, predictive mean matching imputation of a single missing datum proceeds as follows. The imputed value for the j^{th} individual, denoted as $\hat{y}_j$ is imputed as $\hat{y}_j = y_k$ where $(\hat{\mu}_j - \hat{\mu}_k)^2 \leq (\hat{\mu}_j - \hat{\mu}_l)^2$ for every respondent with complete data l, $\hat{\mu}_j$ is the predicted mean of y for the j^{th} individual, and y_k is the observed value of y for the k^{th} respondent. Aside from its simple implementation, another key benefit of this method is that only plausible values for the missing variable will be imputed since the values are transferred from records without missing data. For example, a missing expenditure variable will always be filled in with a value greater than or equal to zero.

The above imputation method can be extended to the case of multivariate item nonresponse. That is, suppose $\mathbf{y}$ is a vector of dimension $(p \times 1)$ with all or some values missing. One could carry out the imputation process described in the preceding paragraph separately for each missing component of $\mathbf{y}$, but this could potentially distort associations among the components of the $\mathbf{y}$ vector. If these associations are important, then a more appropriate strategy would be to define and evaluate a distance metric on the $\mathbf{y}$ vector, match recipients to the "closest" donor, and then transfer the values of the entire $\mathbf{y}$ vector to the recipient as the imputed values.

This imputation strategy may be implemented as follows. The multivariate regression of the $(p \times 1)$ vector $\mathbf{y}$ on a set of covariates, say of dimension $(q \times 1)$, $\mathbf{x}$, yields a $(p \times 1)$ vector of predicted means $\hat{\mu}_j = \hat{\mu}(x_j)$ for every j^{th} individual. We then match each recipient to the closest donor as determined by the Mahalanobis distance function,

$$d^2(j,k) = (\hat{\mu}_j - \hat{\mu}_k)^\intercal \, S_{yx}^{-1} \, (\hat{\mu}_j - \hat{\mu}_k) \tag{1}$$

where S_{yx} is the residual covariance matrix of $\mathbf{y}$ on $\mathbf{x}$. Finally, all of the values of $\mathbf{y}$ for the donor are transferred to the recipient as the imputed values regardless of whether some of the values of $\mathbf{y}$ for are non-missing for the recipient. The MEPS program uses a modified version of this approach which we describe in the next section.

3.2 MEPS' implementation of predictive mean matching

Predictive mean matching is implemented by the MEPS program as follows. First, imputation for *flat fee*, or global fee, events (i.e., events for which one total charge is rendered for a collection of services, such as orthodontia or obstetrical services that encompass a series of health care events) are conducted separately from *simple* events (i.e., health care events covered by a single charge, such as a visit to an office-based physician for an annual physical). In addition, separate imputations of missing payment information are conducted for each event type. We only discuss the imputation of payments for simple MPC-eligible events in this paper.

For each set of simple MPC-eligible events, we define the *recipient* events as those with missing payment information from some or all of the payment sources (These payment sources are identified in Section 2.1.1). In general, the *donor* events are those with a complete set of payment information; however, payment information for an event can still be considered complete even if total charges for the event are missing or were not reported. Furthermore, only MPC-reported events that matched to HC-reported events are included in the set of donors; therefore, unmatched MPC-reported events and HC-reported events that did not match to MPC-reported events were excluded from the donor pool, even if the event had a complete set of payment information. The rationale for only including matched MPC-HC events in the donor pool is two-fold: (1) event-level HC-reported information is used in regression model estimation and (2) MPC charge and payment information are considered to be higher quality.

Imputation begins with estimating the parameters of the regression model given in equation (2) using only the events in the donor pool.

$$y_d = \mathbf{x_d}\beta + \epsilon_\mathbf{d} \tag{2}$$

In equation (2), y_d refers to the square root of the total payment amount of event d in the donor pool, $\mathbf{x_d}$ is a row vector of covariates for event d, and β is a vector of regression coefficients, and ϵ_d is the error term for event d. Recall that total payments are obtained by summing the payment amounts from the individual sources. We denote the estimated regression model as follows.

$$\hat{y}_d = \mathbf{x_d}\hat{\beta} \tag{3}$$

There are four additional points worth noting about equations (2) and (3). First, the dependent variable, y_d, is the square root of total payments for the event. When predictive mean matching was first being explored for utilization in MEPS' imputation procedures, it was determined, through a series of appropriate model diagnostics, that the square root transformation of total payments provided a better model fit (as judged by the R^2-criteria) than both the untransformed version and the logarithmic transformation of the dependent variable. Second, y_d is a univariate quantity. Despite potentially multiple sources of missing payment information in need of imputation, MEPS' implementation of predictive mean matching does not involve a multivariate regression of a $(p \times 1)$ vector $\mathbf{y}$ on a set of covariates $\mathbf{x}$. This is deviation from the method described in Section 3.1 for the imputation of multivariate item nonresponse. Third, as previously mentioned, an event can have complete payment information event if the total charges for the event are missing. When charge information is available for both recipients and donors, it is included in the set of covariates, $\mathbf{x_d}$. Finally, the set of covariates $\mathbf{x_d}$ can be partitioned into three main categories of covariates: (1) *class* variables; (2) indicators for health conditions for the person who had the health care event; and, (3) county characteristics.

The first category of $\mathbf{x_d}$, the *class* variables[2], varies depending on the event type being imputed. The same health condition indicators and county characteristics are incorporated into the regression model regardless of the event type. The *class* variables primarily include indicators for the availability of the source of healthcare coverage (ISOPs), indicators of the source and reported status of payment amounts (SOPs), variables that consolidate the source of payment information (DELTAs), and other characteristics such as indicators for whether the person had an X-ray or laboratory tests, whether surgery was performed, doctor specialty group, and the reason for and length of the hospital stay.

Using the estimated regression coefficients, $\hat{\beta}$, from equation (3), we obtain the predicted mean square root of total payments for all recipient events as follows.

$$\hat{y}_r = \mathbf{x_r}\hat{\beta} \tag{4}$$

In equation (4), $\hat{y}_r$ is the predicted square root of total payments for recipient event r and $\mathbf{x_r}$ contains the values of the covariates in $\mathbf{x_d}$ for recipient event r.

Then, using the predicted mean values, $\hat{y}_d$ and $\hat{y}_r$, we define and evaluate a distance function for each pair of recipient and donor events in source class h. This distance function is given as follows.

$$\delta_{hrd} = |\hat{y}_{hr} - \hat{y}_{hd}| \tag{5}$$

In equation (5), we include the subscript h to denote the payment source class for the recipient and donor events. By payment source class, we mean a classification system

[2]The naming convention of the *class* variables stems from the prior use of the weighted sequential hot-deck to impute missing payment information. Specifically, these variables were used to form the imputation classes under the former approach.

to distinguish among various combinations of potential payers for the healthcare service event. For example, one payment source class would contain all events of persons with the combination of private insurance, Medicare, and OOP payments as potential payers. This property is included so as to guard against the potential matching of a recipient to a donor with a different set of payment sources. For example, this restriction would prevent matching a donor event of a person with only private insurance coverage (and possibly OOP) to a recipient event of a person with Medicaid coverage. We discuss these payment source classes in the next section.

Finally, the donor for recipient r in payment source class h is the donor k which satisfies the following criteria.

$$\delta_{hrk} = \min_{d \in D_h} \delta_{hrd} \tag{6}$$

In words, the donor is the event with the smallest difference between its estimated square root total payments for the event and the recipient's estimated square root total payments among all donor events in the same payment source class h. Once the donor is identified for the recipient only the missing payment amounts are filled-in by transferring the corresponding payment amounts for the donor event to the recipient. In other words, non-missing, complete payment amounts for the recipient event are preserved. We also note that steps are taken to ensure that a specific donor is not matched to too many recipients as using the same donor too many times may distort distributions and/or attenuate associations in the completed data set.

3.3 Special imputation procedures related to Medicaid beneficiaries

The primary motivation for including payment source classes in the distance function given in equation (5) is to prevent matching a recipient to a donor with different potential payment sources. This is also the method by which missing payment information for Medicaid beneficiaries is imputed. We should note that missing payment information for simple MPC-eligible events of Medicaid beneficiaries is essentially handled no differently than similar type events for persons covered by private insurance or in other insurance coverage situations.

The payment source classes are formed using the ISOP and DELTA variables. Recall that the ISOP variables are indicators for the availability of the source of healthcare coverage to serve as a potential payer for the event while the DELTA variables consolidate payer and payment information. There are ISOP and DELTA variables corresponding to each individual source of payment (see Section 2.1.1 for a listing of the various payment sources). Since this information is based on respondent reports of coverage and payment status, it is possible that respondents may not know or refuse to report this information. Thus, different situations can arise depending on whether the coverage (payer is or is not a potential payer for the event) and payment statuses (paid or not paid) were known or reported for all sources. In general, if the coverage and payment statuses for a payer were unknown or not reported, then the value of the DELTA variable corresponding to that payer was set to missing and payment source classes were only formed using the available information. Consequently, Medicaid payments are only imputed for persons who reported in the HC that Medicaid paid for or was a potential payer for the health care service event.

There are 18 payment source classes, but only four of these pertain to Medicaid beneficiaries. These four are

1. Medicare and Medicaid coverage (no private insurance coverage, may or may not have VA coverage) with out-of-pocket;

2. Medicare and Medicaid coverage (no private insurance coverage, may or may not have VA coverage) with no out-of-pocket;

3. Medicaid coverage only (may or may not have private insurance coverage) with out-of-pocket; and,

4. Medicaid coverage only (may or may not have private insurance coverage) with no out-of-pocket.

A complete listing of all 18 payment source classes can be found in Westat (2013).

4. Medicaid beneficiaries' expenditure information

In this section, we provide descriptive statistics to describe the data source (e.g., household component, medical provider component, or imputation) for payment information for Medicaid beneficiaries for their reported office-based physician visits and hospital care in 2011. All results in this section are weighted to reflect the population of the 2011 US civilian, noninstitutionalized Medicaid population under age 65[3]. There were a total of $9,662$ responding sample persons that reported being covered by Medicaid during any round of data collection in 2011 (i.e., Medicaid was reported as a payer or potential payer for any potential health care service event)[4] and a total of $23,156$ HC-reported physician visits and hospital care event for those persons.

Table 1: Total, Medicaid, and OOP payments (in millions) for Medicaid beneficiaries by event type

	Office-based	Outpatient	Inpatient hospital stays	Emergency rooms
Number of events	19,234	1,608	609	1,705
Weighted number of events (in millions)	130.9	10.9	39.5	11.2
Total payments	18,244	5,528	29,662	4,411
Total Medicaid payments	15,072	4,158	24,432	3,527
Total OOP payments	384	30	38	46

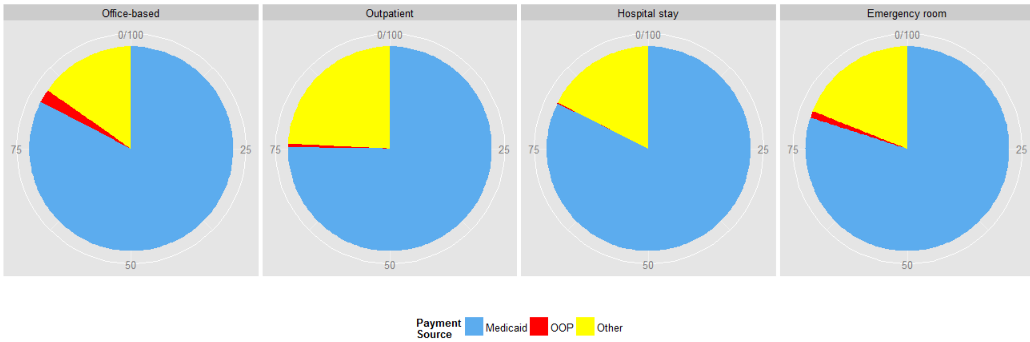
In Table 1, we present the total number of office-based physician visits and hospital care events (broken down by outpatient department visits, inpatient hospital stays, and emergency room visits) for responding sample members who reported Medicaid coverage at some point in 2011. We also show total payments, Medicaid payments, and OOP payments for these events. Using this information in conjunction with Figure 1, we observe that Medicaid pays for over 75% of all expenses for events of each health care service type for Medicaid beneficiaries under age 65. Furthermore, OOP amounts comprise a small proportion of total payments for these events, ranging from 0.13% for inpatient hospital stays to 2.1% for office-based physician visits, while other various sources pay for the remaining portion.

From Table 2, we observe that a smaller proportion of office-based physician visit events have Medicaid payments coming from the MPC and a larger proportion have imputed Medicaid payments when compared to hospital care events. Specifically, a small

[3]Given the differences in MPC sampling methodology for office-based physician visits and hospital care events and the lack of independence between the sets of events we do not compute tests of statistical significance.

[4]Persons who reported being covered by any private source of health insurance coverage or by Medicare were excluded from this analysis.

Figure 1: Percent of total payments for each event type paid by Medicaid, OOP, and Other sources for Medicaid beneficiaries



proportion of each event type have Medicaid payments reported in the HC (5.4%, 6.3%, 3.5%, and 2.9% for office-based physician visits, outpatient department visits, inpatient hospital stays, and emergency room visits, respectively). However, these are actually zero-dollar Medicaid payments[5]. For these events, the respondent reported Medicaid as being a potential payer, but then also reported that Medicaid did not pay for any of the services received during that visit (see Table 3).

Table 2: Percent of Medicaid beneficiaries' events' OOP and Medicaid payments obtained from each source by event type

Source for	Office-based		Outpatient		Hospital stays		Emergency room	
estimation	OOP	Medicaid	OOP	Medicaid	OOP	Medicaid	OOP	Medicaid
HC	30.21	5.36	25.51	6.23	22.87	3.53	20.25	2.89
MPC	66.87	59.77	72.57	66.58	73.88	62.34	77.88	70.82
Imputation	2.92	34.88	1.92	27.19	3.25	34.14	1.87	26.28

In Table 3, we present the percent of each of total OOP and Medicaid payments for Medicaid beneficiaries' office-based physician visits and hospital care events that are obtained from the HC or MPC or are imputed. We observe that for each of the four event types, no Medicaid payments contained on the Public Use Files are obtained from the HC.

Table 3: Percent of all OOP and Medicaid payments for Medicaid beneficiaries obtained from each source by event type

Source for	Office-based		Outpatient		Hospital stays		Emergency room	
estimation	OOP	Medicaid	OOP	Medicaid	OOP	Medicaid	OOP	Medicaid
HC	28.45	0.00	15.96	0.00	30.84	0.00	24.34	0.00
MPC	42.10	48.86	64.54	74.36	28.07	61.98	49.28	64.37
Imputation	29.45	51.14	19.50	25.64	41.10	38.02	26.38	35.63

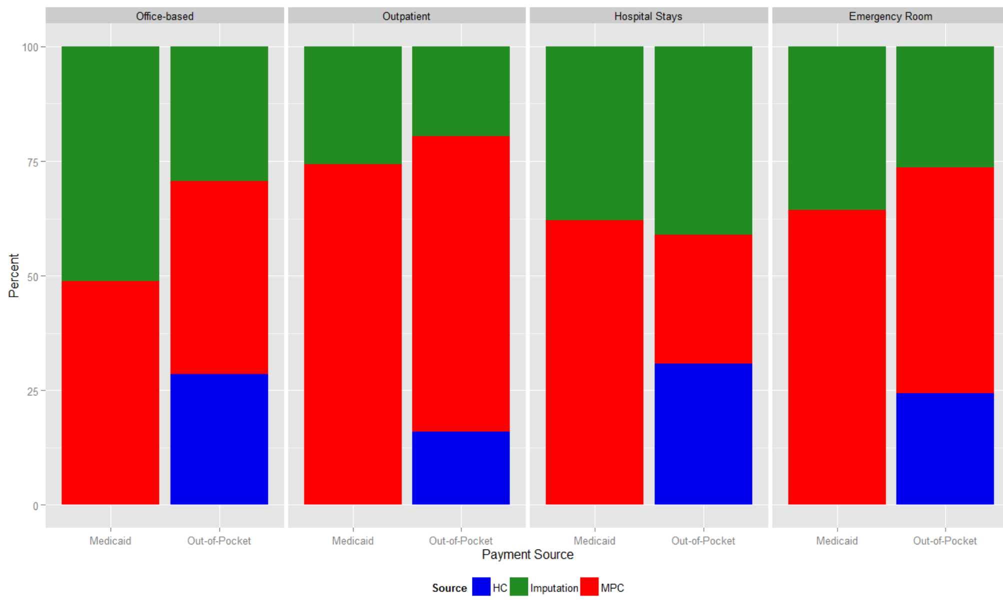
This finding reflects how MEPS' payment data for Medicaid beneficiaries are processed and estimated. In Section 1, we pointed out that even if a respondent reports a non-zero

[5]We also point out that none of these events are covered under a flat fee arrangement since flat fee events were excluded from this analysis.

Medicaid payment for the event, it is ignored and replaced with either medical provider data or through imputation.

To compare how the source of data for Medicaid payments varies among the event types, we display graphically in Figure 2 the descriptive statistics from Table 3. As expected, there is a higher proportion of Medicaid payments for office-based physician visits (51.14%) being imputed than for each of the hospital care event types (25.6%, 38.02%, and 35.6% for outpatient department visits, inpatient hospital stays, and emergency room visits, respectively). This is a consequence of the MPC subsampling procedures for these event types. Recall that even though data on household reported physician visits for Medicaid beneficiaries have a higher probability of being supplemented with data from medical providers, the MPC still only employs a subsample of these events due to budgetary constraints (with the subsampling probability varying by year). Thus, we would expect a higher percent of payments to be imputed for office-based physician visits than for hospital care events because hospital care events are always included in the MPC (provided that the respondent signs an authorization form giving permission to collect these data from the medical provider).

Figure 2: Percent of Medicaid and OOP payments for Medicaid beneficiaries obtained from each data source, by event type



Finally, from Tables 2 and 3, we can evaluate how the data sources for Medicaid payments compare to the data sources for OOP payments for the same events. Across all event types a greater proportion of events' OOP payments and total OOP payments are obtained via the HC than total Medicaid payments and events' Medicaid payments. This is attributable to the fact that respondents are more likely to have knowledge of payments from this source because they are being paid by themselves or their family members. In contrast, as previously mentioned, Medicaid transactions occur between the state Medicaid agency and the medical provider/plan, so the respondent is not expected to have knowledge of these payments.

5. Discussion

In this paper, we provided an overview of the imputation procedures implemented in the MEPS program to fill-in missing payment information for Medicaid beneficiaries' physician visits and hospital care events. We also provided descriptive statistics to convey the amount of OOP and Medicaid payment information that is being reported by the household in the HC, collected via medical providers in the MPC, or completed via imputation methods, namely predictive mean matching. We found that for Medicaid beneficiaries' office-based physician visits and hospital care events', payments by Medicaid are obtained from the medical provider or filled-in via imputation methods. This is a consequence of the post-data collection processing procedures utilized by MEPS. Specifically, the MEPS program ignores non-zero dollar HC-reported Medicaid payments for office-based physician visits and hospital care events and replaces them with medical provider data or fills them in via imputation.

Given our review of the imputation procedures implemented by the MEPS program, we have identified a few additional areas that may warrant further study so as to improve the overall estimation processes employed in MEPS. First, MEPS' data processing and production procedures are anchored in household reporting. This means that events only have the potential of being included in the Public Use Files if they are reported by the household respondent in the HC. Furthermore, not only are MEPS' respondents asked to report all health care service events for the reference period, but they are also asked to identify all potential sources of payment, or coverage, for those events. If the respondent fails to report Medicaid as a payer or potential payer for a particular event when, in fact, it is, then Medicaid payments will not be associated with the event. This is the case even if a medical provider reports payments by Medicaid for the event. Instead, these payments are included in a series of variables designed to reconcile any inconsistencies in reports on health insurance coverage and payment sources for the health care service event. To address this concern, research could focus on survey design strategies to improve the quality of retrospective recall and instrument design modifications to promote the consistency of reported information across sections of the CAPI instrument. For any of the proposed strategies or modifications, this research would also include an investigation of the trade-offs between improved data quality and respondent burden and survey costs.

A second area of possible research is to explore the extent to which, if any, augmenting the donor pool with unmatched MPC-reported events and HC-reported events that did not match to MPC-reported events might enhance the imputation procedures. Recall that these types of events are excluded from the donor pool even if they had a complete set of payment information. Excluding these events has at least two implications. First, the size of the donor pool is smaller than it would otherwise be; thus, the probability that events will serve as donors multiple times is increased. Using a donor too many times may attenuate associations or distort distributions in the data. This may be particularly problematic for Medicaid beneficiaries' events since imputation is done within payment source class and there are likely fewer of these events, in general, due to only about 20% of the civilian noninstitutionalized population, under 65, covered by public sources (in 2011), when compared to those covered by private insurance, about 60% (Agency for Healthcare Research and Quality 2013d). Secondly, MPC data are often viewed as a "gold standard" data source since they are extracted from both the medical providers' administrative records and the associated billing departments. The inclusion of unmatched MPC events in the imputation procedures, particularly the estimation of equation (2), might yield better predictions of event-level total payments; thereby, improving the matching of donors to recipients.

Another area of possible research is to explore alternative regression model specifications to appropriately account for the multivariate nature of the payment amounts for the various payment sources. Recall that in Section 3.2, the dependent variable in equation (2) is the square root of total payments for the event. This is a univariate quantity despite the fact that the imputation problem is is inherently multivariate. Future research could focus on the extent to which, if any, explicitly modeling $\mathbf{y}$ (as a vector) would enhance the imputation procedures. This would also require research on exploring alternative distance functions to accommodate the multivariate vector of predicted values.

Finally, in this paper, we only discussed imputation methods for missing payment information from simple MPC-eligible events. Imputation of Medicaid payments for simple non-MPC eligible events, i.e., dental, other medical, home health by paid independent providers, and non-doctor providers, follows different processes since MEPS does not collect medical provider data for these event types. In particular, if the event has a Medicaid status, then similar to simple MPC-eligible events any household-reported non-zero dollar Medicaid payments are ignored, but then an expected total payment (sum across all sources) is imputed for the event, and the Medicaid payment amount is set equal to the expected total payment less the sum of other reported payments. Here, the expected total payment is the total charge (which itself is often imputed) multiplied by a predetermined discount rate. This discount rate is the ratio of total payments to total charges and varies by event type. Effective use of this type of imputation procedure is predicated on using accurate and time-relevant discount rates; thus, the MEPS program should continuously monitor and evaluate these discount rates and update as necessary. One option to accomplish this is to analyze matched MEPS-MPC-Medicaid claims records and other sources of administrative data to estimate these discount rates (Mirel and Gonzalez, 2014).

Disclaimer and acknowledgments

The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred. The authors would also to thank Bob Baskin, Fran Chevarley, Sadeq Chowdhury, Joel Cohen, Steve Cohen, Chris Cox, Ralph DiGaetano, Huseyin Goskel, D.E.B. Potter, Steve Machlin, Marc Roemer, Diana Wobus, Lap-Ming Wun, Nina Verevkina, and Sam Zuvekas for their insightful and helpful contributions to this paper.

REFERENCES

Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services (2013a). *MEPS-HC Response Rates by Panel*. http://meps.ahrq.gov/mepsweb/survey_comp/hc_response_rate.jsp (visited August 11, 2014).

Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services (2013b). *MEPS-MPC Response Rates, Hospitals*. http://meps.ahrq.gov/mepsweb/survey_comp/mpc_response_rate.jsp?provider_id=11&year=2011&Submit1=Search (visited August 11, 2014).

Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services (2013c). *MEPS-MPC Response Rates, Office-based physicians*. http://meps.ahrq.gov/mepsweb/survey_comp/mpc_response_rate.jsp?provider_id=12&year=9999&Submit1=Search (visited August 11, 2014).

Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services (2013d). *Table 2.a Health insurance coverage of the civilian noninstitutionalized population under age 65: Percent by type of coverage and selected population characteristics, United States, 2011*. http://meps.ahrq.gov/mepsweb/data_stats/summ_tables/hc/hlth_insr/2011/t2a_b11.htm (visited August 11, 2014).

Biemer, P. B. and Lyber, L. E. (2003). *Introduction to Survey Quality*. John Wiley & Sons, Inc. Hoboken, New Jersey.

Centers for Medicare and Medicaid Services, U.S. Department of Health and Human Services (2014). *Centers for Medicare & Medicaid Services Medicaid Eligibility*. http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Eligibility/Eligibility.html (visited August 11, 2014).

Cox, B. G. (1980). The Weighted Sequential Hot Deck Imputation Procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 721–6.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, **64**, 1183–1210.

Groves, R. M., Fowler, J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004). *Survey Methodology*. New York: Wiley.

The Henry J. Kaiser Family Foundation (2014). *June 10, 2014 - Status of State Action on the Medicaid Expansion Decision, 2014*. http://kff.org/health-reform/state-indicator/state-activity-around-expanding-medicaid-under-the-affordable-care-act/ (visited August 11, 2014).

Little, R. J. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business and Economic Statistics*, **6**(3), 287–96.

Machlin, S. R., Chowdhury, S. R., Ezzati-Rice, T., DiGaetano, R., Goksel, H., Wun, L. -M., Yu, W., and Kashihara, D. (2010). *Estimation Procedures for the Medical Expenditure Panel Survey Household Component*. Methodology Report #24. September 2010. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr24/mr24.pdf

Machlin, S. R. and Dougherty, D. D. *Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey*. Methodology Report #19. March 2007. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.gov./mepsweb/data_files/publications/mr19/mr19.pdf

Mirel, L. B. and Machlin, S. R. (2013). Enhancing the Medical Expenditure Panel Survey through Data Linkages. *Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference.*

Mirel, L. B. and Gonzalez, J. M. (2014). Enhancing the Medical Expenditure Panel Survey Data on Medicaid Beneficiaries through Data Linkage. *Forthcoming, Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Westat (2013). Predictive Mean Matching Imputation Method Applied to the Expenditure Imputation of the MPC Event Types for the Full Year 2011 Data. *Internal Memo*.