WORKING PAPER #20001

June 2020

Undercounting of Healthcare Utilization in the Medical Expenditure Panel Survey

Emily Mitchell, PhD, Jerrod Anderson, PhD, and Adam Biener PhD

Table of Contents

1 Introduction

Household-reported medical events in the Medical Expenditure Panel Survey (MEPS) may be undercounted. Previous research suggests that approximately 20 percent of ambulatory visits among Medicare beneficiaries are not reported (Zuvekas & Olin, 2009a) and that this underreporting of utilization by household respondents is primarily due to recall error (Zuvekas, 2011).

Although households are encouraged to keep records of their healthcare visits, MEPS respondents do not uniformly rely on memory aids such as calendars or explanation of benefits forms. When memory aids are not used, respondents are less likely to recall all of the medical events in a round of interviewing. Undercounting of events can also vary by event type and interview round. For example, inpatient hospitalizations for MEPS respondents with Medicare coverage tend to be accurately reported (considering Medicare claims data as benchmark), while emergency department and office-based visits are commonly undercounted (Zuvekas & Olin, 2009a). In addition, interview rounds with longer reference periods are more susceptible to underreporting due to recall error, while those in shorter reference periods are more susceptible to overreporting due to telescoping (recalling events that occurred just outside of the true reference period, e.g. reporting a January 4 event as occurring on December 31). Inaccurate counts of events can have implications for estimates of medical expenditures (Zuvekas & Olin, 2009b) or behavioral analysis of MEPS respondents, though miscounting does not necessarily imply there will be bias in associations across variables (Hill, Zuvekas, & Zodet, 2011).

This paper explores whether person characteristics or aspects of the interview such as use of memory aids are correlated with misreporting among respondents. Multivariate analyses are used to isolate the characteristics associated with misreporting while controlling for other covariates. In addition, the paper assesses whether machine learning algorithms can accurately predict whether a household was likely to overreport or underreport events, as well as to predict the actual number of events, as reported in the MEPS Medical Provider Component (MPC). All analyses are based on hospital inpatient (IP) stays, emergency room (ER) visits, office-based medical provider visits (OB/MVs), and outpatient (OP) visits from MEPS 2014–2016 data. Both household and MPC data are collected for these event types.

Objectives:

1. **Who is misreporting?** Identify *general characteristics of households* that are likely to have undercounted events and *key variables* associated with undercounting. Identifying key characteristics associated with undercounting can help determine whether certain groups of respondents are more likely to undercount. Characteristics associated with undercounting can help improve fielding of the MEPS Household Component by flagging households that may benefit from a more experienced interviewer. Additionally, behavioral analyses of MEPS may be biased when failing to adjust for these characteristics. To accomplish this goal, a logistic regression model is estimated to identify variables that may be associated with undercounting.

2. **How severe is the misreporting?** Quantify the *degree of miscounting* by predicting the true number of events for each person. Accurately predicting the "true" number of events (based on provider-reported event counts) may help correct underestimates of utilization due to undercounting. To

quantify the degree of miscounting, machine learning algorithms are used to predict the difference in provider- and household-reported events.

2 Survey Background

The MEPS Household Component (HC) collects data on healthcare utilization and expenditures from a nationally representative sample of the U.S. civilian noninstitutionalized population. Each participating household is interviewed five times (in five rounds) over the course of approximately two and a half years to collect information covering two consecutive calendar years. Trained interviewers utilize computer-assisted personal interview (CAPI) technology to conduct the interview. Usually, a single respondent answers the survey questions about demographics, health conditions, and healthcare visits and expenditures for each person in the household. MEPS asks that this respondent be the person who is the most knowledgeable about the health and healthcare of all persons in the household.

During each interview, households are asked permission to contact the healthcare providers they visited in order to obtain more accurate expenditure data. At the end of each calendar year, their providers are contacted during the MEPS-MPC survey and asked about all healthcare events for the patient for the entire year. Providers that are included in the MEPS-MPC survey include hospitals, physicians, home health agencies, and pharmacies. Dentists and nonphysician providers (e.g., chiropractors, acupuncturists) are not eligible for the MEPS-MPC survey.

Providers are only eligible to be contacted if the household first grants permission by signing an authorization form. Due to budget constraints, only a sample of eligible providers are contacted, with a focus on collecting data from events likely to have high expenditures, as well as events for households that are more likely to have incomplete expenditure data. For instance, all providers associated with hospital-based events are sampled for MEPS-MPC data collection, while only a percentage of office-based physician practices are sampled. However, 100 percent of providers are sampled for persons on Medicaid, as Medicaid recipients rarely receive billing records for medical events.

While MEPS-HC data are collected by interviewing households five times over the course of two and a half years, MEPS-MPC data are collected from providers once after the end of a calendar year. For hospital events, expenditure information is usually abstracted from billing and medical records. For office-based providers, about half of the events are abstracted; for the remaining events, the billing offices are interviewed over the phone to gather the necessary information. Figure 1 provides a schematic of the interview timeline for panels 19 and 20, where a panel refers to the entire collection of households that enter the survey in the same year.

After the HC and MPC data collection is complete, the MPC events are matched with HC events, and MPC-reported expenditures are transferred to the HC events. This matching process is probabilistic, as exact correspondence between HC and MPC events is not guaranteed. In many cases, the number of events from MPC data collection exceed the number of events reported by the household, although the reverse can also occur.

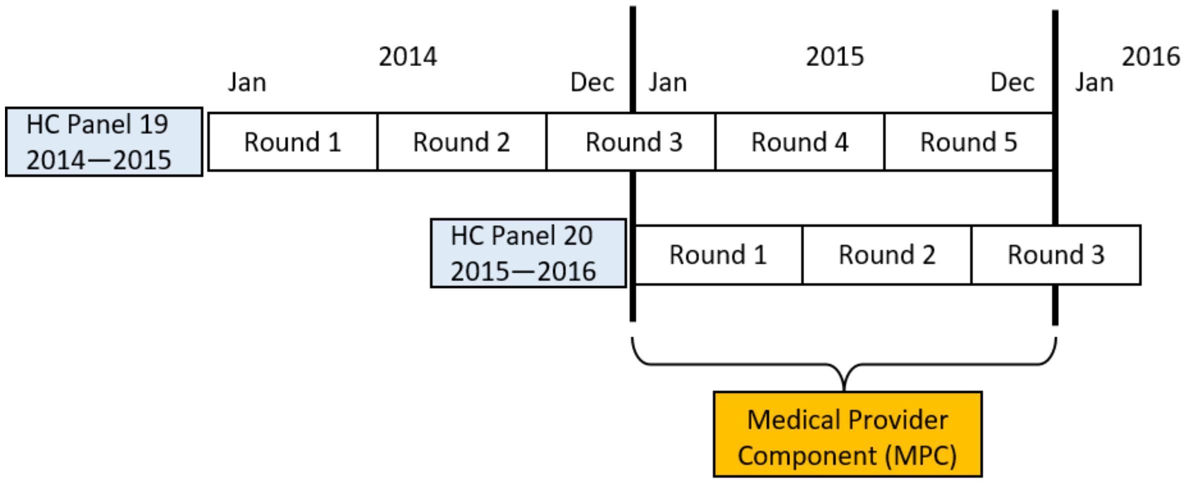

Figure 1. Interview timeline for household panels 19 and 20, and MPC data collection in 2015 for panels 19 and 20.

While the main purpose in collecting provider data in the MPC is to obtain more accurate expenditure information, MPC data are used in this evaluation to assess the magnitude of undercounting or overcounting healthcare events for MEPS participants. Because MPC-reported data are less susceptible to recall error than HC-reported data, MPC data can be used as a starting point to determine the true number of events for a particular person in a given round. Figure 2 provides a schematic of how apparent undercounts or overcounts of events can occur during the data collection process.

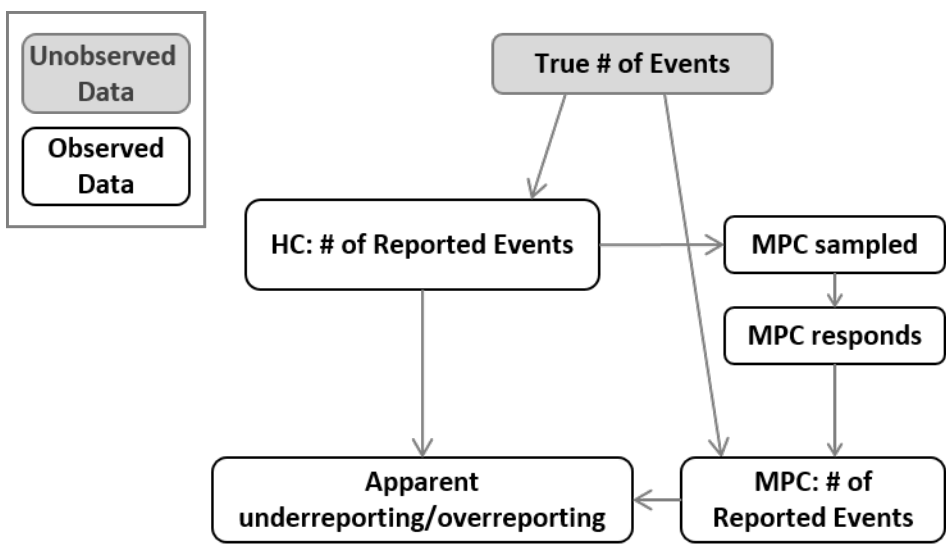

Figure 2. Process from true number of events to apparent overcount or undercount

Underreporting occurs when the household reports fewer events than they actually had, and overreporting occurs when the household reports more events than they actually had. Common reasons for underreporting by the household respondent include recall error or confusion about the provider, whereby events are attributed to an incorrect provider. A simple example would be a household member who saw Dr. Jones twice in a round but reported one visit with Dr. Jones and a second visit with Dr. Smith, resulting in underreporting of one visit with Dr. Jones.

Event miscounts may also occur depending on *how* events are reported, even if the event count is not underreported. One source of miscounting is differences in how households and providers categorize events. For instance, a household may report an office-based medical visit, whereas the provider might report the same event as an outpatient visit. This event-typing inconsistency is common in MEPS due to the similarity between office-based and outpatient visits from the perspective of respondents. Independent doctors' offices located on or near hospital campuses may appear to respondents to be hospital outpatient departments. Similarly, visits to satellite clinics or physician practices owned by hospital systems may be billed as hospital outpatient department visits, but respondents may perceive these to be office-based visits. In some cases, if a stand-alone physician practice is bought out by a hospital system (an occurrence that has become increasingly common), visits to the same provider might be billed as office-based visits prior to the buyout and outpatient visits afterwards, even if the location and doctors do not change. The household may continue to report these events as office-based visits, even if the MPC counts them as outpatient visits.

In other cases, the event type reported by providers may be merely an artifact of billing and reimbursement systems. For example, instead of billing for a single IP stay, hospitals will sometimes bill for multiple consecutive outpatient stays, commonly referred to as "observation" stays. Presently, Medicare will only allow hospital stays that span two midnights or more to be billed as an IP stay; otherwise they must be billed as hospital outpatient observations stays. To the household respondent, the hospital visit may be indistinguishable from a single IP stay, whereas MPC billing records can indicate multiple outpatient stays. This inconsistency in event-type reporting will manifest as an apparent undercount of the household's healthcare utilization.

Overcounting of household events can occur due to telescoping, attributing the same event to multiple reference periods, or incorrectly recalling the number of repeat visits. In rare cases, abstraction errors can occur in the MPC events, resulting in an incorrect number of events (i.e., underreporting or overreporting on the MPC side). This can result in an apparent miscount, where it appears that the household is misreporting the number of events, when in fact the MPC data does not reflect the correct number of events.

While MPC data provide an excellent starting point, determining which MPC events to count, and how to match them to HC data, can prove difficult. Due to the complex nature of the HC and MPC data collection and processing in MEPS, directly comparing the number of events in the HC and MPC data for a specific provider for a particular person (a "person-provider pair") is not as straightforward as it may seem. The next section provides details about how events are counted, and how underreporting and overreporting are defined for this analysis.

3 Datasets

The datasets used for this analysis contain variables needed to count HC and MPC events (*medical_match* dataset), identify potential errors in MPC data collection ("MPC files"), predict a person's true number of events ("Public use files"), and predict whether a person's events are likely to be overcounted or undercounted ("Public use files," "Para-data," and "Geo-code files"). All datasets are internal to the Agency for Healthcare Research and Quality, with the exception of the public use files, which are available for download on the MEPS website. Data years 2014–2016 were used in this analysis. While 2017 was available, design changes in the survey collection instrument resulted in inconsistent variable names across different rounds in 2017. For simplicity, data from 2017 were not included.

Medical match

The medical match dataset (*medical_match.sas7bdat*) is the main dataset used to predict overreporting and underreporting. The purpose of this dataset is to combine HC and MPC events into a single dataset in order to "match" them by identifying MPC event data that is most likely to correspond to each HC-reported event. Direct matching is problematic because exact dates and provider names can be prone to error or entirely missing from HC data. Thus, probabilistic matching is used to select the MPC event most similar to a particular HC event. After matching, expenditure information from the MPC (which is considered more reliable than that of the HC) replaces any HC-provided expenditure information for that event.

For this analysis of miscounting, instead of using this dataset to "match" one HC event to one MPC event, it is used to count the number of HC- and MPC-reported events in a particular round, to compare HC- and MPC-reported utilization counts. In the *medical_match* dataset, each observation is a reported medical event from either the provider or the household. The variable FILE is used to distinguish between events reported by the household respondent (FILE = 1) and those reported by providers (FILE = 2). Additional variables from this dataset include event details (e.g., round, event type, date(s), lab test indicator), person and provider identification numbers, and information on the matching process. Starting in 2016, MPC events were replicated when the HC reported more events within a person-provider pair. This method imputed missing expenditure information from similar events within the same person-provider pair, rather than potentially imputing from other persons or providers. For purposes of the current investigation, these replicated MPC events are removed from the dataset before counting MPC events.

MPC files

Event-level MPC files include data from office-based providers (*obdf1_all*) that is collected mainly through telephone interviews with billing offices, as well as data from hospitals and institutions (*hspf1_all*), which is mainly gathered through abstraction of billing records. Variables in these MPC datasets are used to identify potential errors in MPC data collection. One potential MPC data collection error occurs when lump-sum payments are made. In these cases, a billing record might show a single payment that covers multiple medical visits. During abstraction, it is possible that the single line item would be confused for a single visit, rather than a single payment for multiple visits. The variable PVT_LUMP_SUM_PAYMENTS in the MPC files can help identify these potentially erroneous MPC cases.

As in the medical match file, starting in 2016, these MPC files include replicated MPC events, resulting in a slightly different identification (ID) structure that must be accommodated when merging datasets.

Public use files

Person-level variables from the full-year consolidated (FYC) file contain sociodemographic information that may be predictive of the actual number of events as well as the proclivity for underreporting or overreporting. The FYC file is a person-level file that contains annual summary variables (e.g., AGELAST, SEX) as well as round-specific variables (e.g., in-scope status variables INSCOP31, INSCOP42, and INSCOP53). Relevant variables in this file were converted from wide to long format in order to produce a dataset with one row per person-round.

Para-data

MEPS interview para-data are collected at the interview level, where each observation is a dwelling unit (the physical structure where respondents reside), and each dataset represents a different round. These datasets contain information on the interview process, such as whether the interviewer was new or experienced, and how long the interview lasted. Variables from the National Health Interview Survey (NHIS) data are also included, such as partial/complete status.[1]

Geo-code files

Geo-coded data are collected at the round level for each DUPERSID. The geo-code dataset includes information on the location of the interview that allows for coding metropolitan statistical area (MSA) status and Census division.

4 Defining Inclusion Criteria

The most straightforward method to compare HC and MPC event counts is to compare the number of events reported for a person for a particular provider (a "person-provider pair") to the number of events collected from that provider directly, for the same person-provider pair. For example, if John Smith reports going to Dr. Acula three times in 2014, but Dr. Acula's office reports that John Smith visited four times that year, then John Smith has underreported his utilization with Dr. Acula by a count of one event. However, quantifying utilization in this manner is not so straightforward, as detailed in the following sections.

4.1 Match-eligible events

In some cases, it may be useful to exclude specific HC or MPC events when comparing utilization. Generally, many of the reasons to exclude an event from counting are the same as the reasons that an event would be excluded from expenditure matching.

On the expenditure matching dataset (*medical_match*), a variable indicating matching eligibility (MATCH_ELIGIBLE) is created during the matching process. On the MPC side, an event is ineligible for matching if it is a duplicate event, a rollup (more discussion on rollups is below), has insufficient charge

[1] MEPS participants are a subset of responders to the NHIS. The NHIS variable of "partial/complete" status indicates whether a person completed the NHIS ("complete") or started the survey but did not complete it ("partial"). This variable is used as a potential indicator of the quality of MEPS data.

and payment data (more discussion below), or falls outside of the reference year. On the HC side, an event is match-ineligible if it falls outside the reference year, has no corresponding MPC data for that provider, or if all events for the provider are ineligible for matching.

Match eligibility is the baseline criterion for counting HC and MPC events. However, there are additional reasons for excluding events from utilization counts. The next section considers some of these potential exclusion criteria. The remaining events are referred to as “count-eligible.”

4.2 Count-eligible events

In the current matching process, there is an emphasis on matching as many HC and MPC events as possible in order to use the most MPC data, and subsequently, the most accurate expenditure data. When comparing utilization, however, it is important to be more discerning when it comes to including events for counting and to consider the distribution of HC vs. MPC utilization based on provider definitions, payment type, and event characteristics. In the following tables, HC and MPC counts are defined at the person-provider (PERSID-GID2, a combination of person identifier and provider identifier) level for 2014–2016 data (more details on GID2 are in the following section). Any events that are not match-eligible are removed before counting, but additional exclusion criteria are not implemented (i.e., enumerated events in the tables are not hierarchically excluded). The goal for this data-cleaning process is to preserve the original data as much as possible, while excluding events or person-provider pairs that are likely a result of MPC data abstraction errors.

4.2.1 Provider Identification: GID vs. GID2

To count MPC events properly for each person-provider pair, a reliable and consistent identifier for providers is needed. When households first report that they visited a provider, the provider is assigned an identifier (PROVID). PROVID might be different for different rounds. For example, suppose the aforementioned Dr. Acula works in a physician group called Phantastic Physicians. In his round 1 interview, John Smith may report that he visited Dr. Acula for a routine checkup. In round 2, however, he might instead report that he went to Phantastic Physicians for another visit. In that case, Dr. Acula might be assigned one PROVID, while Phantastic Physicians is assigned another. In the subsequent MPC data collection step, these two unique PROVIDs would be collapsed into a single contact group ID (GID) to avoid contacting Phantastic Physicians multiple times during the MPC data collection process.

On the HC data, events without corresponding MPC data will have a missing contact group ID (GID = .). MPC data for certain providers might not be collected either because the HC did not give permission to contact the provider, the provider was not sampled for MPC data collection (this mainly applies to office-based providers), the provider did not respond for that person, or insufficient data were collected from the provider for that patient.

During the matching process, when MPC events are linked to HC events to transfer expenditure data, an additional provider ID is created by collapsing contact groups (GID2). Contact groups are collapsed “if the provider names or other contact information are similar between any two providers that belong to different contact groups."[1]

The purpose of creating this collapsed contact group ID is to combine providers with similar names or identifying information that could create misclassification errors if the HC respondent attributes an event to the wrong provider. Continuing with the previous scenario, suppose John Smith goes to Phantastic Physicians as well as Phenomenal Physicians multiple times during the year. He may inadvertently misreport which practice he visited for a particular event, especially if both practices are on the same street or otherwise similar. In that case, creating a collapsed contact group that includes both practices will improve the chances that the correct MPC event is matched to the corresponding HC event. Figure 3 shows a schematic of the provider identification process from PROVID to GID to GID2.

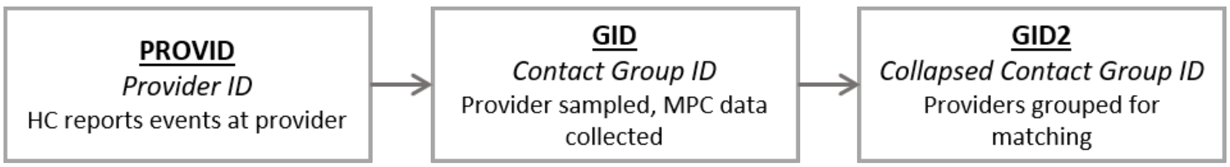

Figure 3. Provider IDs, from HC data collection to MPC data collection to matching

While collapsed contact group IDs (GID2s) help improve the matching process, the creation of GID2 might be too liberal for counting events by person-provider pair. For instance, to continue the previous example where John Smith reported visiting Phantastic Physicians as well as Phenomenal Physicians, Phantastic Physicians is selected for MPC data collection (and is assigned a GID), but Phenomenal Physicians is not (and GID is not assigned). Due to the similarity of the names of the practices, John Smith's HC events for both practices might be collapsed into the same GID2 to improve matching for expenditures. However, because MPC events from Phenomenal Physicians were not obtained, the number of HC-reported events for the collapsed GID2 will likely be greater than the number of MPC-reported events, resulting in false overreporting. In this scenario, using GID2 as the provider ID to compare HC and MPC utilization counts can be misleading, because the MPC events from Phenomenal Physicians were not collected.

The current investigation aims to collapse providers appropriately while avoiding superfluous collapsing of providers. To do this, it is important to consider the construction of the collapsed provider ID (GID2), based on the contact group ID (GID). Table 1 shows the comparison between HC and MPC match-eligible events, based on person-provider pairs defined by GID2, for different types of collapsing. The "large providers" group includes those that typically have a centralized billing system. In these cases, the collapsing of contact groups is based more on historical data than on similarity between provider names and addresses. Groups with a "single GID" are those with simple one-to-one correspondence between GID and GID2 (no collapsing of GID occurs). The group labeled "has multiple (non-missing) GIDs" includes collapsed provider groups (GID2s) that contain multiple providers (based on GID). The group labeled "has missing GIDs" includes collapsed provider IDs (GID2s) that contain household events from

[1] Medical Expenditure Panel Survey – Medical Provider Component (MEPS-MPC). 2016 MPC Plan for Matching MPC to HC Events. Deliverable OP4-28 (April 24, 2017).

providers that do not have an associated GID (i.e., no MPC data was collected based on these contact group IDs).

Table 1. Distribution of reporting by collapsed provider ID makeup (GID2), 2014–2016

GID vs. GID2, by year	# of PERSID-GID2 pairs (%)		
	HC < MPC	HC = MPC	HC > MPC
1 Large provider			
2014	179 (62)	51 (18)	59 (20)
2015	172 (63)	39 (14)	61 (22)
2016	188 (59)	45 (14)	87 (27)
2 Single GID			
2014	7,955 (39)	9,876 (48)	2,652 (13)
2015	6,969 (41)	7,682 (45)	2,291 (14)
2016	7,575 (42)	7,998 (45)	2,299 (13)
3 Has missing GIDs			
2014	1,113 (44)	526 (21)	903 (36)
2015	1,229 (44)	511 (18)	1,036 (37)
2016	1,389 (47)	521 (18)	1,051 (35)
4 Has multiple (non-missing) GIDs			
2014	152 (63)	37 (15)	54 (22)
2015	217 (61)	71 (20)	70 (20)
2016	306 (63)	83 (17)	98 (20)
Overall distribution			
2014	9,399 (40)	10,490 (45)	3,668 (16)
2015	8,587 (42)	8,303 (41)	3,458 (17)
2016	9,458 (44)	8,647 (40)	3,535 (16)

Table 1 shows that the distribution of HC vs. MPC utilization is sensitive to the creation of GID2. In particular, among GID2s that contain missing GIDs (“3 Has Missing GIDs”), 35–37 percent of person-provider pairs have more HC events than MPC events, compared with just 16–17% in the overall distributions. This anomaly supports the assumption that the collapsing of contact groups may be overly liberal, particularly when the constituent contact groups (GIDs) contain HC events with no corresponding MPC data (i.e., GID = .). Thus, these groups should be removed from the analysis comparing utilization counts. Moreover, it is crucial not to merely remove the HC *events* with missing GIDs, as some collapsing among these groups could in fact be appropriate, and removing only HC events could skew the distribution in the opposite direction. Thus, entire person-provider pairs are removed where GID2 contains any missing GIDs (with the exception of large providers).

4.2.2 Rollups

Rollups are groups of MPC events that could be construed as a single event. For instance, if the MPC reports consecutive or overlapping IP stays solely for billing purposes, this probably should have been considered only a single event with respect to counting utilization. Table 2 lists the nine MPC rollup groups and their eligibility for matching.

The groups that are ineligible for matching (and subsequently, ineligible for counting) are consecutive or overlapping IP or institutional care stays (rollup groups #1 and #2); multiple outpatient (OP) events embedded in an IP stay (rollup group #6); and multiple OB/MVs, OP visits, or emergency room (ER) visits that occur on the same day with the same provider (rollup group #9). Each of these situations is likely to be either an artifact of MPC billing practices or an error in MPC data abstraction. In these cases, these events should be considered a single event.

Table 2. List of rollup groups and match-eligible status

MPC Rollup Group	Match-Eligible	Count-Eligible
1 = Multiple consecutive or overlapping IP stays	No	No
2 = Multiple consecutive institutional care stays	No	No
3 = ER events that occur on first day of (or day before) IP stay	Yes	Yes
4 = Multiple OB/MVs and OP visits that occur on same day	Yes	Yes
5 = Consecutive IP stays with 1-day gap	Yes	Yes
6 = Multiple OP events embedded in IP stay	No	No
7 = OP events that occur on first day of (or day before) IP stay	**Yes**	**No**
8 = ER events that occur on same day as one or more OP visits	**Yes**	**No**
9 = Multiple OB/MVs, OP visits, ER visits on same day with the same provider	No	No

Rollup groups #7 and #8 are considered eligible for matching, but it may not be advisable to count all the events in these groups as unique events. OP events that are on the first day of an IP stay (rollup group #7) are highly unlikely to have been reported by a respondent, and are more likely an artifact of the MPC billing process, where OP events are billed separately during an IP stay even though they occurred during the stay. Similarly, OP events that occur on the same day as an ER event (rollup group #8) are often lab or radiology charges billed as an OP event even though they likely occurred during the ER visit. Thus, the OP events in rollup groups #7 and #8 are removed from count eligibility.

Rollup groups that remain eligible for matching and counting include ER events that occur on the first day of an IP stay (rollup group #3), multiple OB/MVs and OP visits that occur on the same day (rollup group #4), and consecutive IP stays with a 1-day gap (rollup group #5). Each of these events is counted as a distinct MPC visit.

4.2.3 Insufficient charge and payment data

MPC events that are match-ineligible due to insufficient charge and payment data (MATCH_ELIGIBLE = 17) are excluded from matching, and thus are excluded from being counted in this analysis. Person-provider pairs with any events that have a MATCH_ELIGIBLE value of 17 are removed, including events that would otherwise have been counted (i.e., where MATCH_ELIGIBLE = 1). The entire person-provider pair (PERSID-GID2) is excluded because the inability to gather sufficient payment data for some MPC events at a provider could reflect errors in reporting events in general for that provider.

4.2.4 Global fees, lump-sum payments, and repeat visits

Global fees, lump-sum payments, and repeat-visit situations might all be susceptible to event undercounting in MPC data collection.

A global fee/flat fee is a fixed dollar amount charged for a package of services (e.g., a fee covering prenatal care, postnatal care, and labor and delivery). For these events, MEPS assumes payments are associated with the first event covered by the fee (referred to as the “stem” event), while the subsequent “leaf” events are assigned $0 in expenditures. A lump-sum payment is easily confused with global fees, because, as is the case with global fees, a single payment is made for multiple visits. However, unlike global fees, nonzero charges and expenditures exist for these multiple visits; the lump-sum payment is merely a convenient method to pay for several visits at once. For both of these types of event groups, it is possible that only a single event would be recorded in lieu of recording all actual visits, since only one payment may be recorded.

Collecting data on repeat visits might cause similar problems, depending on how the billing records are created for repeat visits. It is possible that a single visit is recorded when in fact multiple identical visits actually occurred. For instance, if a person has weekly medical visits for allergy shots, the provider might send a single bill to the patient that covers a month of visits. During MPC data collection, this single bill might be misconstrued as a single visit, resulting in undercounting of utilization on the MPC side. However, indicators of repeat visits in the MPC are rare (encompassing fewer than 60 events yearly, for 2014–2016), since the requirements for treating MPC events as repeat visits are strict. Thus, there is not a reliable method to identify these potential errors in the MPC data.

In the 2014–2016 data, none of the person-provider pairs that reported having global fees or repeat visits had only one MPC event. For lump sums, however, there were over 200 person-provider pairs in each year in which the provider data contained a lump-sum payment but reported only a single MPC event. It is possible that these lump-sum payments could have spanned data collection years so that the single MPC event count for the year is actually correct. However, because these events could indicate an MPC error, person-provider pairs are removed in which the provider reports lump sums but has only one MPC event.

4.2.5 Capitation

Capitation in the context of the MPC refers to how the practice is paid by the insurer, and it is usually a fixed per-person per-month payment for however many patients in that practice are enrolled with that particular capitated plan. In general, providers in a capitated system maintain sufficient encounter records for MPC data abstraction, and there is no evidence that capitated providers are more susceptible to event miscounting in the MPC data collection process. Thus, no events were removed based on capitation status.

4.2.6 ER-IP link

Frequently, when a person is admitted to the hospital for an IP stay, he or she is admitted from the ER. In these cases, ideally two distinct events, one ER and one IP, would be reported. However, identifying ER-IP linked events as two distinct events can be confusing for household respondents, household interviewers, and MPC data abstractors. Thus, this working paper examines situations where the hospital provider reports an IP stay that was not immediately preceded by an ER visit, which could indicate that the ER event was missing in MPC data abstraction.

Table 3 shows that MPC data with IP events that are not linked to an ER event ("1 No IP-ER linked events") has a similar reporting distribution to the overall data. In addition, when the MPC reports at least one set of IP-ER linked events, there are likely to be fewer HC than MPC events (63–67 percent), which is probably due to undercounting on the HC side, rather than an error in MPC data collection. A proportion of these household undercounts are likely attributable to an idiosyncrasy of the CAPI system, rather than respondent recall error. Specifically, whenever a household reported an IP stay, the interviewer was prompted to ask whether the stay was preceded by an ER visit, and if so, to create the ER event and link it to the IP stay. However, the process of creating the ER event was not automated, and many of these reported ER events were ultimately not recorded. This systemic error was addressed during an upgrade of the CAPI system in 2018. For purposes of the current investigation in predicting miscounted household events, no cases are removed from being count-eligible based on an ER-IP link.

Table 3. Distribution of reporting by ER visits linked to IP stays, 2014–2016

	# of PERSID-GID2 pairs (%)		
ER-IP link (MPC), by year	**HC < MPC**	**HC = MPC**	**HC > MPC**
0 No IP events			
2014	8,232 (39)	9,829 (46)	3,274 (15)
2015	7,598 (41)	7,706 (42)	3,111 (17)
2016	8,353 (43)	8,085 (41)	3,203 (16)
1 No IP-ER linked events			
2014	624 (45)	490 (35)	275 (20)
2015	537 (44)	446 (37)	236 (19)
2016	585 (48)	420 (34)	219 (18)
2 At least one IP-ER linked event			
2014	543 (65)	171 (21)	119 (14)
2015	452 (63)	151 (21)	111 (16)
2016	520 (67)	142 (18)	113 (15)
Overall distribution			
2014	9,399 (40)	10,490 (45)	3,668 (16)
2015	8,587 (42)	8,303 (41)	3,458 (17)
2016	9,458 (44)	8,647 (40)	3,535 (16)

4.2.7 Lab tests

When collecting data on medical visits, households are asked whether a particular event included a lab test. Similarly, the MPC data collection process records whether lab tests were performed. On the HC side, the concern is that a household may underreport lab test events. For instance, suppose Jane Smith has a prenatal visit with her obstetrician-gynecologist (OB/GYN) and is then instructed to go to LabWorks, a separate lab-testing facility, for blood work on the same day. Jane may mistakenly report only her visit to the OB/GYN (perhaps reporting a lab test during that visit) but fail to report the separate visit to LabWorks, resulting in Jane underreporting her visit to LabWorks. On the MPC side, the concern is that during MPC abstraction, a lab test may be misconstrued as a separate visit, even though the patient only visited the provider once. Suppose instead that Jane Smith's OB/GYN drew blood during Jane's visit, rather than sending Jane to LabWorks. Afterward, the OB/GYN sends Jane's blood sample to a lab-testing facility, where it is tested, and services are billed on the following day. In this scenario, the MPC abstractor may mistake the date of the lab test as a separate medical visit, when in fact only a single medical visit should be recorded.

Table 4 shows the comparison between HC and MPC utilization counts, based on the reporting of lab tests within each person-provider pair. For the majority of pairs, neither the HC nor MPC report an event with lab tests ("0 No lab tests"). When both the HC and MPC report lab tests, there tend to be slightly more pairs for which the HC counts are fewer than MPC counts (49–54 percent) compared with the overall distribution (40–44 percent). This category likely contains some lab tests that were incorrectly reported as separate events on the MPC side. However, it is difficult to distinguish MPC lab events that are true "visits" from those that were incorrectly recorded as separate visits due to confusion during abstraction. For instance, if a patient has blood drawn during an OP visit, but the sample is processed the next day, the date of the lab test has a distinct date from the date of the visit in the medical records, and it may appear to be a separate, stand-alone event. Instead of attempting to exclude erroneous MPC lab events, this investigation instead uses the reporting of lab tests as a predictor variable in subsequent analyses of miscounting.

Table 4. Distribution of reporting by MPC lab test makeup, 2014–2016

	# of PERSID-GID2 pairs (%)		
Lab tests reported, by year	HC < MPC	HC = MPC	HC > MPC
0 No lab tests			
2014	3,489 (31)	6,301 (56)	1,445 (13)
2015	2,696 (32)	4,567 (54)	1,193 (14)
2016	2,918 (33)	4,790 (54)	1,199 (13)
1 MPC lab tests only			
2014	2,153 (60)	1,114 (31)	342 (9)
2015	2,057 (61)	993 (29)	346 (10)
2016	2,267 (64)	946 (27)	331 (9)
2 HC lab tests only			
2014	839 (31)	1,087 (40)	794 (29)
2015	875 (32)	989 (37)	838 (31)
2016	937 (32)	1,106 (37)	917 (31)
3 HC and MPC lab tests			
2014	2,918 (49)	1,988 (33)	1,087 (18)
2015	2,959 (51)	1,754 (30)	1,081 (19)
2016	3,336 (54)	1,805 (29)	1,088 (17)
Overall distribution			
2014	9,399 (40)	10,490 (45)	3,668 (16)
2015	8,587 (42)	8,303 (41)	3,458 (17)

Lab tests reported, by year	# of PERSID-GID2 pairs (%)		
	HC < MPC	HC = MPC	HC > MPC
2016	9,458 (44)	8,647 (40)	3,535 (16)

5 Analysis

5.1 Data preparation

The analysis described in this paper is based on the *medical_match* datasets, which contain HC and MPC events for office-based visits, outpatient visits, ER visits, and IP stays. HC-reported office-based visits to nonphysicians (e.g., chiropractors, physical therapists) are ineligible for MPC data collection and thus are not included in this analysis. In addition, the dataset for this analysis does not include MPC data for providers that were not sampled for MPC data collection (a subset of office-based providers), or that were sampled but did not respond to the data collection effort.

To tease out person-provider pairs that may have erroneous MPC data (as discussed in detail in the previous section), the following person-provider pairs (including both HC and MPC events) were removed from the analytical dataset:

- Person-provider pairs (PERSID-GID2) with missing GID
- Person-provider pairs (PERSID-GID2) with any events that are ineligible for matching due to insufficient charge and payment data (MATCH_ELIGIBLE = 17)
- Person-provider pairs (PERSID-GID2) where the MPC reported lump-sum payments but recorded only one MPC event

The dataset also excluded any events that are match-ineligible (MATCH_ELIGIBLE ≠ 1), including events that fell outside of the reference year, as well as events for persons that are not in the public use file (due to survey attrition), had a non-positive sample weight, or did not appear in the MPC file (i.e., no MPC data were collected for that person). In addition, events were removed that fell during a round for which the person was not in scope during the entire round, because MPC events could be collected during this period, but HC events would not have been collected. Other events that were removed include those for persons with missing para-data information, MPC events that were replicated for matching (2016 only), and MPC child events in rollup groups #7 and #8 (specifically, OP events that occurred on the first day of an IP stay, or on the same day as an ER event).

Table 5 displays the number of HC and MPC events at each stage of data cleaning, after removing ineligible events from the *medical_match* dataset.

Table 5. Data cleaning process, 2014–2016

	Number of events remaining					
	2014		2015		2016	
	HC	MPC	HC	MPC	HC	MPC
1. Medical_match dataset	162,371	108,544	163,515	97,734	157,427	154,864
2. Remove replicates (2016 only)	162,371	108,544	163,515	97,734	157,427	110,583
3. Remove match-ineligible events (MATCH_ELIGIBLE ≠ 1)	81,407	99,815	70,388	89,484	75,842	99,787
4. Remove child events in rollup groups #7 and #8	81,407	99,341	70,388	88,946	75,842	99,163
5. Remove pairs with any missing GIDs	61,660	78,667	50,209	66,929	53,283	73,768
6. Remove pairs with any insufficient charge and payment data (MATCH_ELIGIBLE = 17)	61,155	78,323	49,867	66,591	52,743	73,113
7. Remove pairs with lump-sum payments but only one MPC event	60,858	78,105	49,473	66,325	52,434	72,881
8. Remove rounds where person not in scope entire round	59,938	76,700	48,696	64,825	51,475	71,225
9. Remove persons with missing para-data	59,933	76,687	48,696	64,825	51,474	71,220

5.2 Person-rounds

Once all the count-ineligible events were removed from the analytical dataset, the number of HC and MPC events were then counted by collapsing to each person-round. All events reported for a person in a round were summed to get the number of HC-reported events. All events reported by providers for that person in that round were summed to get the number of MPC-reported events. Table 6 gives an example of the event-level *medical_match* dataset prior to collapsing, and table 7 shows an example of the same subset of events on the person-round-level dataset after collapsing (data are fabricated and for expository purposes only).

Table 6. Event-level "medical match" file

DUPERSID	Round	In Scope?	GID2	Record ID	Count-Eligible?	File
20031112	3	Yes	210	1001	1	HC
20031112	3	Yes	210	1002	1	HC
20031112	3	Yes	210	1003	1	MPC
20031112	4	Yes	210	1004	1	MPC
20031112	4	Yes	210	1005	1	MPC
10029877	1	Yes	83	9010	1	HC
10029877	1	Yes	83	9011	1	MPC
10029877	1	Yes	410	9012	0	HC
10029877	1	Yes	410	9013	1	MPC
10029877	2	No	83	9014	1	HC
10029877	2	No	83	9015	1	MPC
10029877	2	No	410	9015	1	MPC
10029877	3	Yes	83	9016	1	MPC
10029877	3	Yes	83	9017	1	MPC
10029877	3	Yes	83	9018	0	MPC

Notes: Cells with a gray background are removed prior to counting

Table 7. Person-round level analytical file to be used in analysis of HC reporting

DUPERSID	Round	In Scope?	# HC Events	# MPC Events	Report	Difference (MPC – HC)
20031112	3	Yes	2	1	Over	-1
20031112	4	Yes	0	2	Under	2
20031112	5	Yes	0	0	Equal	0
10029877	1	Yes	1	2	Under	1
10029877	3	Yes	0	2	Under	2

Note that for person 20031112, round 5 was added in the person-round-level collapsed data with HC = 0 and MPC = 0, with a report status of "Equal." For person 10029877, round 2 was removed because the person was not in scope during the entire round. Furthermore, one HC event was removed in round 1 and one MPC event was removed in round 3 because those events were not count-eligible.

This analysis used the person-round-level dataset collapsed across providers (table 7) to compare HC- and MPC-reported events for several reasons.

Why collapse across providers?

As mentioned, direct comparison of HC- and MPC-reported events for a single provider is not straightforward, due to potential HC misreporting and issues with MPC data collection. For purposes of this investigation, if a person saw two providers in a round but only one provider was in the MPC, the HC events associated with the provider that was not in the MPC were not counted. This analysis was more concerned about the overall magnitude of HC underreporting or overreporting of utilization, rather than the magnitude of misreporting *at a particular provider*.

Why round-level instead of year-level counts?

Events were counted within a round, instead of for the entire year, for the following reasons:

(1) *Underreporting/overreporting occurs at the interview level*. Keeping rounds separate allowed use of interview-level variables without sacrificing information that would be lost in creating year-level summary variables of interview-level data.

(2) *Round number can affect overreporting/underreporting*. For instance, overreporting due to telescoping may be more likely to occur in round 1 (due to a shorter reference period), while underreporting due to respondent fatigue may be more likely to occur in later rounds (Zuvekas, 2011).

(3) *This approach allowed for 0/1 comparisons.* If a household fails to report events for a particular provider, that provider is not surveyed about the household's medical events. Furthermore, this household underreporting is unobserved, as no data from the provider is obtained. This is the so-called "0/1" problem, where situations in which the household reports no events for a provider cannot be observed, but the provider would have reported one or more events. Thus, in order to observe any MPC events for a provider, the HC must report at least one event for that provider. At the year level, this means that there can be no 0/1 comparison for HC-MPC

event counts. However, using round-level data allows capture of 0/1 underreporting at the round level. For instance, if a household reports an event at a provider in round 1 but fails to report an event at that provider in round 2, “0/1” underreporting will be observed in round 2, because all events for the year are recorded from the provider in the MPC.

5.3 Summary of reporting

This section summarizes the distribution of reporting by event type and round, because these two variables are likely to be correlated with reporting accuracy. For example, IP stays and ER events are rare and salient events for most people, so it is easier for respondents to accurately recall the number of ER visits or IP stays they had during a reference period. On the other hand, regular OB/MV or OP medical visits may be more easily forgotten, particularly if a respondent has frequent or recurring events.

Interview round may influence underreporting or overreporting because longer reference periods can be associated with a reduced ability to recall medical events. In addition, respondents may succumb to “reporting fatigue” and begin to underreport in the later rounds of the survey as they learn that reporting fewer events will result in a shorter (and less burdensome) interview.

Total HC and MPC events, as well as the overall distribution of reporting status, are displayed in table 8 by event type and year. Person-rounds in which both the HC and MPC report no events for a particular event type are shown in a separate column. Percentages do not include this special case of “equal-reporting.” “Ambulatory” events combine OP and OB/MV events in an effort to mitigate some of the inconsistency in HC and MPC reporting of these two event types, as discussed previously. “All Events” includes IP, ER, OP, and OB/MV event types.

Table 8. Total number of HC and MPC events and reporting distribution by event type, 2014–2016

			Reporting status (person-rounds)			
Event Type	**# HC Events**	**# MPC events**	**Under N (%)**	**Equal N (%)**	**0-0 N**	**Over N (%)**
All Events						
2014	59,933	76,687	10,347 (43)	9,850 (41)	12,059	3,753 (16)
2015	48,696	64,825	10,172 (44)	9,278 (40)	11,928	3,811 (16)
2016	51,474	71,220	11,173 (44)	10,061 (40)	12,274	3,941 (16)
IP						
2014	1,592	1,530	290 (19)	880 (57)	34,457	382 (25)
2015	1,479	1,335	211 (15)	828 (59)	33,795	355 (25)
2016	1,445	1,302	205 (15)	819 (59)	36,066	359 (26)
ER						
2014	4,061	5,889	1,613 (36)	2,293 (51)	31,518	585 (13)
2015	3,707	5,230	1,495 (36)	2,134 (51)	31,031	529 (13)
2016	3,700	5,223	1,457 (35)	2,164 (52)	33,323	505 (12)
Ambulatory						
2014	54,280	69,268	9,818 (45)	8,522 (39)	13,948	3,721 (17)
2015	43,510	58,260	9,673 (45)	8,044 (37)	13,719	3,753 (17)
2016	46,329	64,695	10,688 (45)	8,962 (38)	13,943	3,856 (16)

			Reporting status (person-rounds)			
Event Type	**# HC Events**	**# MPC events**	**Under N (%)**	**Equal N (%)**	**0-0 N**	**Over N (%)**
OP						
2014	7,248	21,126	4,818 (70)	1,431 (21)	29,117	643 (9)
2015	6,023	19,590	4,582 (70)	1,373 (21)	28,668	566 (9)
2016	6,160	21,173	4,972 (72)	1,407 (20)	30,540	530 (8)
OB/MV						
2014	47,032	48,142	7,156 (37)	7,513 (38)	16,416	4,924 (25)
2015	37,487	38,670	6,972 (37)	6,921 (37)	16,304	4,992 (26)
2016	40,169	43,522	7,782 (37)	7,841 (37)	16,516	5,310 (25)

As shown in the table, for 43–44 percent of person-rounds in 2014–2016, household events were undercounted in that round. In addition, for around 16 percent of person-rounds the number of household events was overcounted. This distribution differs by event type. For instance, the percentage of equal-reports for IP stays (57–59 percent) and ER events (51–52 percent) are higher than the overall percentage of equal-reports (40–41 percent). As mentioned, this is likely due to the salience of these types of events, as well as the relative rarity of their occurrence, making it more straightforward for households to report these events accurately.

Additionally, the number of HC-reported OP events is far lower than the number of MPC-reported OP events. This finding is similar to the results from Zuvekas and Olin (2009a) on Medicare claims data, in which the researchers report a similar discrepancy between household- and provider-reported OP events. One potential cause for this difference is that some of the MPC-reported outpatient events might be lab or radiology tests, in which the household never physically visited the facility, but the MPC recorded a lab test as an "event" because it produced a bill. In these cases, the misreporting is actually on the MPC side, because these should not be considered true events. Another potential cause of this discrepancy is inconsistency in event-type reporting by the household and providers. As discussed earlier, the same visit might be reported as an OB/MV event by the household, but as an OP event by the provider. This event-typing difference results in an apparent overcounting of household OB/MV events and an apparent undercounting of household OP events, which is reflected in the relatively high percentage of OB/MV overcounts (25–26 percent) in conjunction with the high percentage of OP undercounts (70–72 percent). As mentioned previously, this event-type confusion is not uncommon, especially considering the increasing expansion of hospital systems that buy out stand-alone physician practices, thus converting what would otherwise be an OB/MV event into an OP event.

Another interesting phenomenon is that households reported more IP stays than providers reported. This could be due to households reporting zero-night hospital stays as IP events, while the provider listed them as ER or OP events. On the other hand, a household may report a multiple-night IP stay, whereas the MPC will categorize the same event as multiple OP visits for billing purposes. As a follow-up to the 2018 CAPI redesign, starting with 2019 data collection, zero-night hospital stays can no longer be reported by the household, but rather must be categorized as ER or OP during the interview.

To mitigate some of this event-typing discrepancy between the HC and MPC in this analysis, event types are combined in the overall prediction model. Analyses were also run on ambulatory events (OP medical visits and OB/MVs) separately, as event-typing discrepancy is particularly prominent between these two event types.

Table 9 shows the distribution of reporting split by panel, round, and year. Note that overreporting is fairly stable across the rounds (14–17 percent), but underreporting tends to be higher for round 5, as well as round 3 in a panel's first year in the survey.

Table 9. Total number of HC and MPC events, and distribution of reporting, by panel and round, 2014–2016

					Reporting status (person-rounds)			
Year	Panel	Round	# HC Events	# MPC Events	Under-reports N (%)	Equal-reports N (%)	0-0 reports N	Over-reports N (%)
2014	18	3	6,354	7,467	1,063 (34)	1,562 (49)	2,903	538 (17)
2014	18	4	13,766	17,394	2,003 (43)	1,950 (42)	1,459	664 (14)
2014	18	5	10,215	14,138	2,086 (50)	1,480 (35)	1,895	631 (15)
2014	19	1	7,741	9,166	1,280 (37)	1,621 (47)	2,426	576 (17)
2014	19	2	13,249	16,758	1,949 (42)	1,946 (42)	1,319	709 (15)
2014	19	3	8,608	11,764	1,966 (51)	1,291 (33)	2,057	635 (16)
2015	19	3	4,359	5,201	917 (35)	1,265 (48)	2,890	459 (17)
2015	19	4	11,144	14,544	1,917 (44)	1,797 (41)	1,136	692 (16)
2015	19	5	8,060	10,996	1,882 (49)	1,344 (35)	1,729	618 (16)
2015	20	1	5,998	7,961	1,323 (38)	1,606 (46)	2,650	580 (17)
2015	20	2	12,233	15,906	2,103 (42)	2,032 (41)	1,208	826 (17)
2015	20	3	6,902	10,217	2,030 (52)	1,234 (32)	2,315	636 (16)
2016	20	3	5,042	6,263	1,032 (33)	1,548 (50)	3,113	514 (17)
2016	20	4	12,131	16,374	2,156 (43)	2,013 (41)	1,253	791 (16)
2016	20	5	8,929	12,953	2,144 (49)	1,519 (35)	1,899	679 (16)
2016	21	1	5,936	7,852	1,379 (39)	1,602 (45)	2,672	590 (17)
2016	21	2	12,238	17,022	2,310 (45)	2,014 (40)	1,157	773 (15)
2016	21	3	7,198	10,756	2,152 (52)	1,365 (33)	2,180	594 (14)

5.4 Logistic regression

The main predictor variable of interest is the number of HC-reported events ("total_events_hc" or "X"). The main outcomes of interest include the following:

- Underreporting/overreporting status
- Magnitude of underreporting/overreporting

Additional predictors are described in table 13 in the appendix. Variables that begin with the prefix "re_" indicate collapsed versions of variables. Health variables that were coded as "inapplicable" were assigned a value of "no" (e.g., priority condition variables for children and PREGNT for males and children). *Any_ref, any_DK,* and *any_NA* represent whether responses for any of the following variables were "refused," "don't Know," or "NA" [not applicable], respectively: pregnancy, priority conditions, [instrumental] activities of daily living (IADL) indicators, marital status, health status, education, and employment. Activities of daily living (ADL) and IADL for rounds 2 and 4 were imputed to be the values in the previous round (questions associated with these variables are only asked during interview rounds 1, 3, and 5).

In the first part of the analyses, a generalized multinomial logistic regression model was run to determine which predictors are associated with higher underreporting or overreporting, for all event types combined.

Table 14 in the appendix shows the coefficient estimates and standard errors of the predictors in the model (run in SAS version 9.4). The level of statistical significance is denoted by asterisks (* = 0.05, ** = 0.01, *** = 0.001).

Misreporting (both underreporting and overreporting) was more likely to occur for people with more events (total_events_hc), lower income (POVCAT = near poor, poor, or negative), public insurance (PUBX = Yes), and for blacks or Hispanics (RACETHX = 1, 3). Underreporting was more common among households with more people (AGE16_64, AGEOV64), when the reference period was longer (REFDAYS) or the interview time shorter (re_INTVTIME), when households were reluctant to respond (re_RELRESP = "Reluctant" or re_NUM_CONT was higher), or when the household's NHIS was incomplete (re_NHIS_COMP = "Partial complete"). Underreporting was also more likely to occur for older persons (AGEX), for persons with a priority condition (re_CANCERDX, re_DIABDX, re_HIBPDX, re_PREGNT, re_STRKDK) or poor health (re_MNHLTH or re_RTHLTH = "fair or poor [or missing]"), and for those with less education (re_education = "High School" or "Less than high school"). Underreporting was more common in round 5, when the reference person was reporting for a child or grandchild (re_REFRLX = "child/grandchild"), and when the household was not on the "calendar path" (CALPATH = "No").

Overreporting was more likely to occur when households were unsure of exact dates of events (higher HC_pct_daymiss), reported duplicate events (HC_pct_duplicates), reported more lab tests (HC_pct_labtest), or reported more repeating events (HC_pct_repeat).

6 Prediction: Machine Learning

In the next portion of the analyses, machine learning models were used to predict reporting status. The algorithms were run on the complete dataset predicting all events (OB/MV, OP, ER, and IP events combined) and each event separately. They were also run on ambulatory events (OB/MV and OP events combined). For the "all events" prediction, additional models were also tested, splitting by adult/children, and separately for person-rounds in which the household reported 0 events (HC = 0) vs. at least one event (HC > 1). Adults and children were split into separate models because adults are more likely to report for themselves, and some of the variables that are predictive of utilization and reporting (such as priority condition variables, employment status, and marital status) are only applicable to adults. Data were split by households that report 0 events in a round in an attempt to better model the "0/1" problem.

For each of these analyses, data were first split into training and test sets using a 70/30 split to avoid over-fitting. For the sub-analyses on the overall event models, the train/test split was performed by blocking on adults/children, year, panel, and round. Person-rounds in which both the household and provider reported no events are retained in the datasets.

6.1 Classification: Identifying households likely to underreport

As above, observations were labeled as under-, equal-, and over-reporters. To see whether reporting status could be predicted accurately, the AdaBoost classification algorithm was used and its performance compared with the multinomial logistic regression. The AdaBoost algorithm builds multiple models in an iterative fashion. For the observations not well predicted in the first model, the AdaBoost algorithm used observations to train a second model. For those observations not well predicted by the second model, a third model was created, etc. Weights were then assigned to each model based on its accuracy. AdaBoost was chosen because it performed better on the combined training data relative to other algorithms. The performance metrics used were the algorithms' accuracy and the area under the receiver operator characteristic (ROC) curve. Accuracy measures how well the model that is developed using the training data correctly identifies the reporting status of each observation in the test data. An ROC curve is the relationship between the true positive rate and the false positive rate as the decision threshold for classification varies. The area under the ROC curve (AUC) provides a composite measure of the tradeoff between the true positive rate and the false positive rate. A curve with a greater AUC will have fewer false positives for every true positive prediction. The AUC can be interpreted in this way: when observations are classified into two groups (e.g., over-report vs. not over-report), the AUC is the probability that a random pair of observations (one from each group) will be correctly classified. Because the outcome contains three categories (rather than two), a "one vs. the rest" classification scheme was used in which the reporting status was converted to a binary outcome (i.e., equal vs. under and over, under vs. equal and over, equal vs. under and over), the ROC curves were calculated, and the average of the results was taken.

Figure 4 displays the ROC curves for all event types combined to provide a comparison of how well the AdaBoost and multinomial logistic models performed. The 45-degree line is plotted for convenience. A model whose ROC curve is equal to the 45-degree line is uninformative. The blue, thick dotted lines are the averaged ROC curve. For the multinomial logistic model, a threshold for classification that allowed for a 20 percent false positive rate would lead to a true positive rate of around 60 percent. For the AdaBoost model, a threshold that allowed for a 20 percent false positive rate would lead to a true positive rate of around 70 percent.

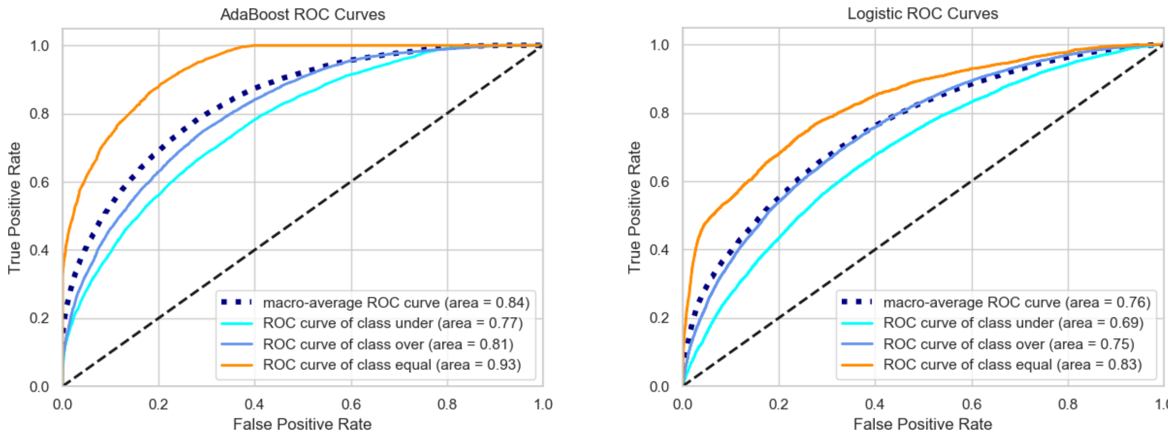

Figure 4. ROC curves for AdaBoost and multinomial logistic models

Notes: The ROC curves of "class under" provide the ROC curve when classifying observations as under-reporters or not under-reporters (i.e., under-report vs. over- and equal-report). The macro-average curve averages all three curves.

In addition to the model with all event types combined, separate models were also run for each event type as well as a model for ambulatory events. The overall model was also run on the following predefined data splits:

1. Adults/children-split
2. Split by HC = 0 and HC > 0

Table 10 provides the accuracy and AUC measures for the separate models. The AdaBoost classification algorithm had higher accuracy in all but two cases (the model for panel 17, round 3 and the model for HC-reported events equal to zero) and equal to or larger AUC in all cases.

Table 10. Accuracy and AUC for the separate models

	AdaBoost		**Multinomial Logistic**	
Model	**Accuracy**	**AUC**	**Accuracy**	**AUC**
IP stays	0.97	0.93	0.98	0.93
ER events	0.94	0.89	0.94	0.88
OP visits	0.86	0.85	0.86	0.81
OB/MVs	0.73	0.83	0.70	0.77
Ambulatory events (OB/MVs + OP visits)	0.71	0.83	0.67	0.76
All events combined	0.71	0.84	0.66	0.76
Adult/child split				
Adults only	0.72	0.85	0.66	0.76
Children only	0.7	0.8	0.67	0.75
HC = 0 / HC > 0				
HC = 0	0.76	0.70	0.76	0.70

Model	AdaBoost		Multinomial Logistic	
	Accuracy	AUC	Accuracy	AUC
HC > 0	0.69	0.85	0.57	0.70

Notes: Higher values of accuracy and AUC indicate better performance, where a value of 1 would indicate perfect predictive ability.

The preferred approach using the AdaBoost algorithm had somewhat superior predictive accuracy and lower rates of expected false positives relative to the multinomial logit. This relatively higher accuracy suggests that observable characteristics contain important information about the reporting behavior of MEPS respondents, potentially even when the true number of medical events is not observable.

6.2 Regression: Estimating the true number of medical events

To estimate the true number of events per person-round using machine learning (ML) algorithms, the outcome of interest was set as the difference in MPC and HC events (total_diff = # of MPC events – # of HC events). After training the ML algorithms on the "train" dataset, the true difference between MPC and HC events on the test set was predicted and added to the original number of HC-reported events. This predicted "true" number of events was then adjusted to conform to the original domain of event counts. Specifically, the predicted value was rounded to the nearest whole number, since partial events are nonsensical. In addition, if a negative number of events was predicted for a person-round, these values were reset to 0.

These adjusted predictions from the test set were used to calculate the R-squared value for each ML algorithm, in order to select the best ML model for prediction (section 6.2.1). Using the selected ML algorithm, the predicted number on the test set were then compared to the number of MPC-reported events (section 6.2.2). All analyses in this section were run in Python 3.

6.2.1 Fitting ML models

Various ML algorithms were tested to predict the difference in all events separately, all events combined, and ambulatory events (OP and OB/MV events combined). For the "overall" (i.e., all events) prediction model, ML algorithms were also run on the aforementioned predefined data splits (adults/children and HC = 0 / HC > 0) to determine whether doing so could result in a better predictive value.

Each ML algorithm was fit on the training set. Predictor variables in the model include variables in table 13, as well as the household-reported event counts (total_events_hc, er_events_hc, etc.). Hyper-parameter tuning was performed using 5-fold cross-validation. Results in the following tables come from applying the prediction algorithms to the test data set.

Table 11 shows the R-squared values for each tested ML algorithm and outcome, where a higher R-squared value indicates better fit, and a value of 1 is perfect fit. "MLR" refers to multiple linear regression. "Elastic Net" is a regularized linear regression model that can protect against over-fitting. "Gradient Boosting" and "Random Forest" are both ensemble methods based on decision trees, but differ slightly in implementation. "Neural Net" is a deep learning algorithm. Additional machine learning algorithms that were similar to those presented in the table were also tested. Results from these additional algorithms are not shown because they did not improve upon the displayed results. For the models on each individual event type, a "multi-output" approach was tested, where the outcomes for ER, OP, OB/MV, and IP event count differences were predicted simultaneously. However, results did not provide a notable improvement over the algorithms that ran each outcome separately; thus, these results are not shown in the table.

Table 11. R-squared values for various ML algorithms and models on test data

Model	MLR	Elastic Net	Gradient Boosting	Random Forest	Neural Net
IP stays	0.58	0.57	0.57	0.58	0.42
ER events	0.30	0.30	0.31	0.30	0.42
OP visits	0.34	0.33	0.39	0.42	-0.05
OB/MV visits	0.23	0.23	0.44	0.48	0.22
Ambulatory events (OB/MVs + OP visits)	0.48	0.48	0.63	0.68	0.48
All events combined	0.50	0.50	0.69	0.69	0.48
Adult/child split	*0.50*	*0.50*	*0.68*	*0.70*	*0.43*
Adults only	0.49	0.49	0.69	0.70	0.42
Children only	0.52	0.52	0.62	0.63	0.47
HC = 0/HC > 0	*0.50*	*0.51*	*0.68*	*0.69*	*0.44*
HC = 0	0.02	0.03	0.02	0.01	-0.60
HC > 0	0.48	0.48	0.69	0.70	0.45

As shown in table 11, in general, Random Forest performed best across all models, and provided a notable improvement over MLR when predicting OP and OB/MV events (both separately and combined as "ambulatory" events). In addition, the predictive accuracy of Random Forest for all events did not improve by forcing predefined splits of the data (adult/children or HC = 0 / HC > 0). In other words, fitting separate models to adult and child datasets did not improve predictive power over predicting on the entire dataset. This is likely due to the nature of Random Forest, which will automatically split the data, for instance, by adults and children, if that split improves predictive ability. None of the algorithms perform well when applied to the sets of person-rounds where the household reports 0 events (R-squared values for all algorithms are less than 0.04).

6.2.2 Predicting true event counts

Following testing of various ML algorithms and predefined data splits, the Random Forest model was chosen on the overall data (with no predefined splits) as the best predictive algorithm. This prediction algorithm was applied to the complete test dataset to get the total number of predicted events and compare those to the "true" number of events (i.e., the number of MPC events). For comparison purposes, the MLR algorithm to predict the true number of events was also applied. Table 12 shows the number of HC-reported events, MPC-reported events, and predicted events, where predicted events are based on the Random Forest and MLR models applied to the test data. The ratio of predicted to MPC-reported events is also shown, where a ratio of 1.00 indicates perfect prediction in the aggregate.

The results in table 12 demonstrate that both MLR and Random Forest performed well in predicting the total number of events for IP stays and ambulatory events. However, both algorithms under-predicted the number of ER events, and the Random Forest algorithm under-predicted OP and OB/MV events. In addition, both algorithms performed poorly when predicting the true number of events for person-rounds where the respondent reported 0 events (HC = 0). Even though the Random Forest algorithm did a better job than MLR at predicting event counts at the person-round level (as seen in Table 11), it did not necessarily improve prediction in the aggregate.

Table 12. Number of HC-reported events, MPC-reported events, and predicted events (based on the Random Forest algorithm on the complete test dataset)

			Predicted (Ratio)	
Model	**HC**	**MPC**	**MLR**	**Random Forest**
IP stays	1,277	1,206	1,240 (1.03)	1,188 (0.99)
ER events	3,432	4,922	3,633 (0.74)	3,738 (0.76)
OP visits	6,232	18,468	18,727 (1.01)	17,057 (0.92)
OB/MVs	37,395	38,770	39,395 (1.02)	36,727 (0.95)
Ambulatory events (OB/MVs + OP visits)	43,627	57,238	58,438 (1.02)	56,814 (0.99)
All events combined	48,336	63,366	64,462 (1.02)	62,977 (0.99)
Adult/child split				
Adults only	40,638	52,207	53,304 (1.02)	51,767 (0.99)
Children only	7,698	11,159	11,158 (1.00)	11,210 (1.00)
HC = 0/HC > 0				
HC = 0	0	7,022	11,429 (1.63)	5,877 (0.84)
HC > 0	48,336	56,344	53,033 (0.94)	57,100 (1.01)

Notes: Ratio = Predicted/MPC-reported events.

7 Summary and Conclusion

This working paper uses the MEPS HC-MPC nexus to assess the degree to which MEPS household medical events are misreported, using medical provider records captured in the MPC as a source of validation data. In this investigation, it was found that from 2014 to 2016, respondent events were underreported for roughly 44 percent of MEPS person-rounds, while about 16 percent of person-rounds were over-reports. This distribution of underreporting or overreporting varies by event type. For instance, IP stays and ER visits are more likely to be accurately reported compared to other event types, which is consistent with prior research showing that MEPS respondents more accurately report high-salience events such as hospital stays. On the other hand, event-typing inconsistency between households and providers contributes to 25–26 percent of overreporting among person-rounds for OB/MVs and 70–72 percent of underreporting for OP visits. These findings can be used to better understand the veracity of MEPS event estimates based on the HC and potentially to allocate survey resources toward individuals more likely to misreport their medical events.

The primary goals of this paper are twofold. First, the investigation aimed to identify characteristics of households likely to have miscounted events. Using a logistic regression model, it was determined that, in general, misreporting was more common among persons with more events as well as people with lower income, public insurance, and for blacks and Hispanics. Secondly, the investigation sought to determine whether machine learning algorithms could be leveraged to predict the likelihood and magnitude of misreporting at the person-round level. Under a classification framework to predict overreporting and underreporting status, the AdaBoost algorithm had a 71 percent accuracy rate, while the multinomial logistic regression model had an accuracy of 66 percent. Both algorithms were better at predicting misreporting for IP stays and ER events, when compared with predictions for OB/MV or OP events. After testing multiple algorithms to predict the true number of events, Random Forest was selected as the most accurate model. However, while the Random Forest algorithm performed best at the person-round level (based on R-squared values), it did not outperform the MLR predictions of aggregate total event counts.

Ideally, machine learning algorithms would be applied to person-rounds that do not have MPC data in order to estimate total household utilization in MEPS. However, the performance of the machine learning algorithms was not sufficient to provide reliable estimates. In addition, initial investigations indicated that the set of person-rounds that were used to train the prediction algorithms is not necessarily generalizable to person-rounds without MPC data; by design, the MPC data collection targets high-cost hospital events, whereas person-rounds without MPC data consist primarily of office-based visits.

A secondary goal of this working paper was to document the steps taken toward identifying "count-eligible" events. Although prior research has used external validation, such as Medicare claims data, to verify reported events for subsets of the MEPS sample, using in-sample MPC data allows for analysis of reporting behavior among MEPS respondents for whom external claims data are not readily available. However, as evidenced in the description of data cleaning and processing in section 4, comparing counts of household-reported events with provider-reported events is not straightforward. Notable issues include potential attribution of events to the wrong provider, difficulties in determining which PROVID to use, and event-typing inconsistency between the HC and MPC. To address some of the difficulties regarding counting events, this investigation elected to collapse events across providers at the person-round level and to run analyses on all event types combined as well as individually. The considerations and issues identified in this paper can be referenced for future work that seeks to use provider-reported utilization counts in quantifying reporting accuracy.

There are limitations to this investigation. First, some unobserved underreporting is likely occurring as a consequence of households omitting all events associated with some providers. If a household fails to report all events for a particular provider, that provider is not surveyed about that household's medical events, and the subsequent underreporting is unobserved. This is the so-called "0/1" problem, where situations where the household reports 0 events for a provider, but the provider would have reported 1 or more events during the year, cannot be observed. In an attempt to capture this deficiency, this investigation evaluated misreporting on a round-by-round basis, where the "0/1" reporting can be captured for a specific round. For instance, if a household reports an event at a provider in round 1, but fails to report an event at that provider in round 2, "0/1" underreporting was observed in round 2, since all events for the year are recorded from the provider. However, this attempt to include "0/1" underreporting does not sufficiently capture the "0/1" problem on an annual scale.

Second, estimates of misreporting rely on the assumption that the MPC is the “gold standard” for event counts and is without error. While this investigation attempted to exclude potentially problematic MPC events before comparing them to HC counts, undoubtedly some MPC observations are themselves measured with error. For instance, there are likely residual errors in the MPC count measure pertaining to how certain events such as lab tests associated with visits are treated. Furthermore, there may be residual errors in households attributing events to the wrong providers, despite efforts to mitigate this potential error by collapsing across providers to the person-round level.

Third, despite the sophistication of the machine learning algorithms used, they are somewhat content-blind. For example, they can predict negative events for a person-round, or partial events, which are not possible values in reality. For this analysis, edits were made on the predicted values to force predicted event counts to be non-negative whole numbers. Future work could consider count-data models that yield positive discretized predictions. Other future work could also provide a more detailed description of the types of respondents for whom the differences between actual and predicted events are largest, and the possible implications these differences have for MEPS expenditure estimates. Future analyses into MEPS miscounting could provide useful insight for statistical, benchmarking, and behavioral analyses using MEPS data, as well as critical information for future iterations of the MEPS instrument.

As mentioned previously, the CAPI instrument used for MEPS household data collection was redesigned in 2018 in an effort to improve user experience for MEPS respondents and thus the quality of the collected data. One of the goals of this redesign was to make it easier for respondents to report recurring events. This change was implemented in part to address some of the concerns of event underreporting in MEPS. The methods outlined in this paper could be used as a baseline for comparing reporting accuracy before and after the CAPI redesign, to help assess the success of the survey redesign in improving the collection of household-reported events.

8 References

Hill, S. C., Zuvekas, S. H., & Zodet, M. W. (2011). Implications of the accuracy of MEPS prescription drug data for health services research. *Inquiry: The Journal of Health Care Organization, Provision, and Financing, 48*(3), 242–259.

Zuvekas, S. H. (2011). The effects of recall length and reporting aids on household reporting of health care events in the Medical Expenditure Panel Survey. *Journal of Economic and Social Measurement*, *36,* 321–343.

Zuvekas, S. H., & Olin, G. (2009a). Validating household reports of health care use in the Medical Expenditure Panel Survey. *Health Services Research, 44*(5), 1679–1699.

Zuvekas, S. H., & Olin, G. (2009b). Accuracy of Medicare expenditures in the Medical Expenditure Panel Survey. *Inquiry, 46*(1), 92–108.

9 Appendix

Abbreviations

ADL	Activities of Daily Living	MPC	Medical Provider Component
ER	Emergency Room	OB/MV	Office-Based Medical Visit
FYC	Full-Year Consolidated	NHIS	National Health Interview Survey
HC	Household Component	OP	Outpatient
[I]ADL	[Instrumental] Activities of Daily Living	PUF	Public Use File
IP/HS	Inpatient/Hospital Stay	RMSE	Root Mean Squared Error
MEPS	Medical Expenditure Panel Survey	SAQ	Self-Administered Questionnaire
ML	Machine Learning	SE	Standard Error

Table 13. Predictor variables and labels by data source and type

Variable Source and Type	Name	Label
FYC PUF		
Survey	PANEL	Panel number
	roundx	Interview round (edited)
	PROXY	Was respondent a proxy
	RESP	1st respondent indicator
	re_REFRLX	Relation to reference person
Demographics	adult	(AGEX >= 18)
	AGEX	Age
	SEX	Sex
	RACETHX	Race/ethnicity
	POVCAT	Poverty category
	re_MARRYX	Marital status
	re_education	Education level
	re_EMPST	Employment status
Insurance Coverage	INSX	Insured
	MCDATX	Medicaid coverage at any time during round
	MCARE	Medicare coverage at any time during round
	PRIVAT	Private coverage at any time during round
	PUBX	Public coverage at any time during round
Health Status	ADLHLP	ADL screener
	IADLHP	IADL screener
	re_CANCERDX	Cancer diagnosis (> 17)
	re_CHOLDX	High cholesterol diagnosis (> 17)
	re_DIABDX	Diabetes diagnosis (> 17)

Variable Source and Type	Name	Label
	re_EMPHDX	Emphysema diagnosis (> 17)
	re_HIBPDX	High blood pressure diagnosis (> 17)
	re_STRKDX	Stroke diagnosis (> 17)
	re_PREGNT	Pregnant during round
	re_MNHLTH	Perceived mental health status
	re_RTHLTH	Perceived health status
SAQ	SAQELIG	Eligibility and completion of SAQ
	re_ADOVER42	SAQ: Can overcome illness without help from a medically trained person
Missings	any_Ref	Any “Refused” values for demo./health vars.
	any_DK	Any “Don’t know” values for demo./health vars.
	any_NA	Any “Not ascertained” values for demo./health vars.
Para-data		
Interview	re_INTVTIME	Length of interview in minutes
	re_TRAINFLAG	Interviewer train flag
	re_INTVLANG	Interview language
	re_INTVTYPE	Mode of interview
	REFDAYS	Reference days in the round
Reluctance to Participate	re_NUM_CONT	Number of total contacts
	re_RELRESP	Reluctance to respond
Household Characteristics	AGE0_15	Total # of persons in household with age 0–15
	AGE16_64	Total # of persons in household with age 16–64
	AGEOV64	Total # of persons in household with age over 64
NHIS variables	re_NHIS_TIME	NHIS interview time
	re_NHIS_COMP	NHIS completion status
Memory Aids	CALPATH	Calendar path
	re_CA_HAVE_REC	Household has records for events
	re_CA_REC_QUAL	Household uses records
	memcat	Summary of memory aids used
Constructed	HC_pct_daymiss	HC visits with day missing (%)
	HC_pct_duplicates	HC duplicate visits (%)
	HC_pct_labtest	HC lab tests (%)
	HC_pct_repeat	HC repeat visits (%)
Geo-codes	censusdiv	Census region

Variable Source and Type	Name	Label
	inMSA	Metropolitan statistical area indicator

Table 14. Multinomial logistic regression results

Variable	Underreporting Estimate (SE)	Overreporting Estimate (SE)
Continuous Variables		
total_events_hc	0.37 (0.01)***	0.42 (0.01)***
AGE0_15 (number of individuals in household under age 15)	0.01 (0.01)	-0.05 (0.01)***
AGE16_64 (number of individuals in household aged 16-64)	0.08 (0.01)***	0.01 (0.02)
AGEOV64 (number of individuals in household aged 65 or older)	0.11 (0.02)***	0.02 (0.03)
AGEX	0.06 (0.01)***	0.04 (0.01)**
HC_pct_daymiss	-0.37 (0.09)***	2.89 (0.08)***
HC_pct_duplicates	-1.26 (0.37)***	5.19 (0.25)***
HC_pct_labtest	-1.25 (0.06)***	2.33 (0.05)***
HC_pct_repeat	-1.46 (0.09)***	2.22 (0.08)***
REFDAYS	0.19 (0.01)***	-0.01 (0.01)
re_NUM_CONT	0.01 (0.002)***	0 (0.003)
Categorical Variables		
CALPATH		
0 No	0.23 (0.07)***	0.25 (0.11)*
1 Yes	REF	REF
INSX		
1 Yes	0.14 (0.04)***	0.06 (0.06)
2 No	REF	REF
MCARE		
1 Yes	-0.1 (0.04)**	-0.09 (0.06)
2 No	REF	REF
MCDATX		
1 Yes	0.09 (0.03)**	0.05 (0.05)
2 No	REF	REF
PANEL		
18	REF	REF
19	0.04 (0.02)	0.2 (0.04)***
20	0.06 (0.02)*	0.27 (0.04)***
21	0.14 (0.03)***	0.29 (0.05)***
POVCAT		
1 Poor/negative (< 100% of PL)	0.25 (0.03)***	0.18 (0.05)***
2 Near poor (100–125% of PL)	0.18 (0.04)***	0.2 (0.06)***
3 Low income (125–200% of PL)	0.09 (0.03)***	0.09 (0.04)*
4 Middle income (200–400% of PL)	0.11 (0.02)***	0.08 (0.03)*
5 High income (> 400% of PL)	REF	REF
PRIVAT		
1 Yes	-0.09 (0.03)***	0 (0.04)
2 No	REF	REF
PROXY		
1 RESPONDENT IS RU MEMBER	REF	REF
2 RESPONDENT IS A PROXY	-0.32 (0.11)**	0.02 (0.16)
PUBX		
1 Yes	0.23 (0.04)***	0.19 (0.06)***

Variable	Underreporting Estimate (SE)	Overreporting Estimate (SE)
2 No	REF	REF
RACETHX		
1 Hispanic	0.17 (0.02)***	0.12 (0.04)**
2 White	REF	REF
3 Black	0.18 (0.02)***	0.23 (0.04)***
4 Asian	0.11 (0.04)**	0.04 (0.06)
5 Other	0.13 (0.04)**	0.1 (0.07)
RESP		
1 YES, FIRST RESPONDENT	REF	REF
2 NO, NOT FIRST RESPONDENT	0.11 (0.02)***	0.02 (0.04)
SAQELIG		
0 Not eligible for SAQ	-0.08 (0.13)	0.04 (0.2)
1 Eligible for SAQ - has data	REF	REF
2 Eligible for SAQ - no data	0.01 (0.06)	-0.05 (0.1)
SEX		
1 MALE	-0.03 (0.02)	0.01 (0.03)
2 FEMALE	REF	REF
adult		
0 child	REF	REF
1 adult	0.27 (0.12)*	0.02 (0.2)
any_DK		
0 No	REF	REF
1 Yes	-0.01 (0.1)	0.05 (0.17)
any_NA		
0 No	REF	REF
1 Yes	-0.39 (0.23)	-0.96 (0.4)*
any_Ref		
0 No	REF	REF
1 Yes	0.3 (0.24)	-0.26 (0.47)
censusdiv		
1 New England	0 (0.04)	0.09 (0.06)
2 Middle Atlantic	0.01 (0.03)	0.11 (0.04)*
3 East North Central	-0.09 (0.03)***	-0.01 (0.04)
4 West North Central	0.01 (0.04)	-0.1 (0.06)
5 South Atlantic	-0.05 (0.02)	-0.07 (0.04)
6 East South Central	-0.01 (0.04)	-0.25 (0.06)***
7 West South Central	0.06 (0.03)*	-0.09 (0.05)
8 Mountain	-0.37 (0.04)***	0.08 (0.06)
9 Pacific	REF	REF
inMSA		
0 No	REF	REF
1 Yes	0.02 (0.02)	-0.05 (0.03)
memcat		
1 Calendar and Records	-0.02 (0.03)	0.37 (0.07)***
3 Only records	-0.05 (0.04)	0.27 (0.07)***
4 No memory aids	REF	REF
re_ADLHLP		
1 Yes	-0.1 (0.05)	-0.1 (0.07)
2 No	REF	REF
re_ADOVER42		
0 Missing	0.2 (0.05)***	0.09 (0.07)

Variable	Underreporting Estimate (SE)	Overreporting Estimate (SE)
1 Disagree	0.1 (0.02)***	0.06 (0.03)
2 Agree or Uncertain	REF	REF
re_CANCERDX		
1 Yes	0.12 (0.03)***	0.05 (0.04)
2 No	REF	REF
re_CA_HAVE_REC		
1 Has MEPS material	REF	REF
2 Has other material	0.07 (0.02)***	-0.04 (0.03)
3 Has no records	0.09 (0.07)	-0.25 (0.11)*
4 Will not use records	0 (0.09)	-0.26 (0.15)
re_CA_REC_QUAL		
1 All visits recorded	REF	REF
2 Most visits recorded	0.1 (0.02)***	0.06 (0.03)*
3 Some visits recorded	0.02 (0.07)	-0.17 (0.12)
4 No records	.	.
5 No events to record	-0.01 (0.05)	-1.12 (0.16)***
re_CHOLDX		
1 Yes	0.03 (0.02)	0.06 (0.03)*
2 No	REF	REF
re_DIABDX		
1 Yes	0.28 (0.02)***	0.02 (0.04)
2 No	REF	REF
re_EMPHDX		
1 Yes	0.11 (0.05)*	0 (0.07)
2 No	REF	REF
re_EMPST		
1 Employed	REF	REF
2 Not employed	0.01 (0.02)	0 (0.04)
re_HIBPDX		
1 Yes	0.13 (0.02)***	0.04 (0.03)
2 No	REF	REF
re_IADLHP		
1 Yes	0.13 (0.04)**	0.12 (0.06)
2 No	REF	REF
re_INTVLANG		
1 In English	REF	REF
2 Not in English	0.06 (0.03)*	0.04 (0.04)
re_INTVTIME		
1 Less than 1 hr	0.17 (0.03)***	-0.29 (0.05)***
2 1 to 1.5 hrs	0.09 (0.02)***	-0.09 (0.04)*
3 1.5 to 2 hrs	REF	REF
4 2 to 3 hrs	-0.08 (0.02)***	0.05 (0.04)
5 3 to 6 hrs	-0.17 (0.03)**	0 (0.04)
6 More than 6 hrs	-0.06 (0.03)	0.01 (0.05)
Missing	0.03 (0.07)	0.07 (0.12)
re_INTVTYPE		
1 In person	REF	REF
2 Other	0.12 (0.04)**	-0.05 (0.08)
re_MARRYX		
1 Married entire round	0.03 (0.02)	-0.03 (0.04)

Variable	Underreporting Estimate (SE)	Overreporting Estimate (SE)
2 Not married	REF	REF
re_MNHLTH		
1 Good or better	REF	REF
2 Fair or Poor (or missing)	0.1 (0.03)***	0.11 (0.04)**
re_NHIS_COMP		
-1 MISSING	0.05 (0.04)	0.05 (0.06)
1 COMPLETE	REF	REF
2 PARTIAL COMPLETE	0.1 (0.02)***	0.09 (0.04)*
re_NHIS_TIME		
1 Less than 1 hr	REF	REF
2 1 to 1.5 hrs	-0.01 (0.02)	0 (0.03)
3 1.5 to 2 hrs	0.01 (0.02)	0 (0.04)
4 More than 2 hrs	-0.02 (0.03)	-0.01 (0.04)
Missing	.	.
re_PREGNT		
1 Yes	0.75 (0.06)***	0.15 (0.09)
2 No	REF	REF
re_REFRLX		
0 Self	REF	REF
1 Spouse/Partner	0.04 (0.03)	0 (0.04)
2 Child/Grandchild	0.16 (0.05)***	-0.04 (0.08)
3 Parent/Grandparent	0 (0.06)	-0.01 (0.09)
4 Other	-0.04 (0.07)	0.03 (0.11)
re_RELRESP		
0 Not reluctant	REF	REF
1 Reluctant (or missing)	0.26 (0.05)***	0.02 (0.08)
re_RTHLTH		
1 Good or better	REF	REF
2 Fair or Poor (or missing)	0.15 (0.02)***	0.02 (0.04)
re_STRKDX		
1 Yes	0.13 (0.04)***	-0.01 (0.05)
2 No	REF	REF
re_TRAINFLAG		
1 Experienced	REF	REF
2 New (or missing)	0.07 (0.02)**	0.1 (0.04)*
re_education		
1 Less than HS	0.09 (0.03)***	-0.03 (0.04)
2 High School	0.08 (0.02)***	-0.01 (0.03)
3 Some college	REF	REF
Missing	0.54 (0.04)***	0.25 (0.07)***
roundX		
1	REF	REF
2	0.17 (0.03)***	0.08 (0.05)
3a	0.14 (0.03)***	0.11 (0.06)
3b	-0.81 (0.04)***	-0.11 (0.06)
4	0.19 (0.03)***	0.13 (0.05)*
5	0.44 (0.03)***	0.25 (0.05)***

Notes: Reference category for the outcome is "Equal-reporting" (including 0-0s). Significance level is denoted by asterisks: * = 0.05, ** = 0.01, *** = 0.001. re_NUM_CONT was imputed to the mode (3) when NUM_CONT was "-9 not ascertained.