Assessing the Quality of Blended Statistics in a Federal Health Survey

Liz Kantor[1], Ann Bisognano[1], Michael Trierweiler[1]
[1]NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603

Abstract
The Medicare Current Beneficiary Survey (MCBS) is a continuous longitudinal survey of a nationally representative sample of the Medicare population, including interviews with facility staff on behalf of beneficiaries living in long-term care facilities. Interviewers and facility staff obtain selected survey items from facility and medical records, many of which are redundant to the administrative data certified facilities regularly report to the Centers for Medicare and Medicaid services (CMS), such as the Minimum Data Set (MDS). To reduce burden, the MCBS instrument was redesigned in 2019 to skip these items if a CMS Certification Number (CCN) can be used in record linkage is reported during the interview.

The CCN, present in about half of cases, is used in a deterministic record linkage protocol during data processing and administrative data are substituted for skipped variables to create a blended data product with survey data. This paper evaluates the quality of these blended statistics when compared to past years' data that are abstracted from records or reported in the survey. We assess the accuracy and completeness of the record linkage protocol, data comparability, and comparative changes in levels of item non-response. Overall, this protocol resulted in a near-complete match rate with largely comparable blended data alongside significant operational gains.

Key Words: Administrative data, survey data, blended statistics, data quality, record linkage, item nonresponse

1. Introduction

The ability to link administrative data to government surveys offers the potential to reduce respondent burden and enhance data quality in those surveys. Administrative data have been used to improve estimates or survey responses through methods such as imputation models, calibration and poststratification, and direct substitution of responses from records identified via either probabilistic or deterministic record linkage (Davern et al., 2019; Lohr and Raghunathan, 2017). In recent years, several large-scale government surveys, including the American Community Survey (ACS), the Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS), the National Health Interview Survey (NHIS), and the Medicare Current Beneficiary Survey (MCBS), have experimented with incorporating some of these techniques (U.S. Census Bureau, 2015; Gordon et al., 2018; Seeskin et al., 2018).

These efforts, among others, show promise for creating effective blended data products for government surveys which incorporate administrative or big data sources while reducing burden. Evaluating their quality and usefulness, however, relies on an understanding of potential pitfalls of these new data sources and their optimal use, through methods such as an extension of the well-known Total Survey Error framework (Groves et al. 2010, Japec et al., 2015). For example, bias may result from incomplete

sampling frames, nonresponse, and the process of data integration (Auerbach et al. 2019). Additionally, the potential for administrative data to be used in conjunction with a survey depends on its accuracy and timeliness in addition to its comparability to the survey data (Seeskin et al. 2018).

The present study focuses on evaluating the quality of a new process to create blended data products for the Facility component of the MCBS after the survey instrument was redesigned to integrate existing administrative data (Mayfield et al. 2019).

1.1 Introduction to the MCBS

The Medicare Current Beneficiary Survey (MCBS) is a continuous, multipurpose survey of a nationally representative sample of the Medicare population, conducted by the Centers for Medicare and Medicaid Services (CMS) through a contract with NORC at the University of Chicago. Since 1991, the survey has served as the leading source of information on the Medicare program and its impact on beneficiaries. The MCBS utilizes a round-based rotating panel design that collects data for beneficiaries over four years for beneficiaries living both in community and facility settings.

The MCBS Facility instrument collects data about beneficiaries residing in long-term care settings such as nursing homes and skilled nursing facilities. One major section of the interview is the Facility Questionnaire (FQ), which collects information about the certification status, bed counts, and services offered where the beneficiary currently resides. Additionally, the Health Status (HS) section is administered at the health assessment level, collecting information on the beneficiary's health status, medical conditions, and functioning at different points in time. It also collects the date and type of health assessment forms in the beneficiary's medical records. The interview is conducted by trained field interviewers with facility staff members rather than the beneficiary themselves. Prior to a redesign of the Facility instrument, information entered into the questionnaire was obtained manually, or abstracted, by either the interviewer or staff from existing facility and medical records for all MCBS beneficiaries in long-term care settings.

1.2 Facility Instrument Redesign

As part of a larger initiative to reduce respondent burden and enhance data quality, NORC conducted research focused on investigating ways to shorten Facility Instrument content by using administrative data and streamlining the remaining content. This investigation uncovered significant portions of the FQ and HS sections of the instrument which aligned with existing administrative data held by CMS. The existing HS section was largely designed to mirror the Long-Term Care Minimum Data Set (MDS) health assessment, and responses to the HS section were already frequently abstracted from MDS assessments conducted for the beneficiary. Several FQ questions were also redundant with facility characteristics that exist in the Certification and Survey Provider Enhanced Reports (CASPER) database. Both administrative data sources store federally mandated quality reporting information.

Medicare- and Medicaid-certified long-term care facilities, which account for approximately half of facilities where MCBS beneficiaries reside, are required to regularly report MDS and CASPER administrative data to CMS. All long-term care facilities certified to provide care under Medicare and Medicaid have a CMS Certification Number (CCN) that uniquely identifies the facility and is present in both the MDS and CASPER datasets.

After a feasibility study, the MCBS Facility instrument was redesigned in 2019 to capture the CCN, which facilitated skipping items that overlap with MDS and CASPER administrative data (Carnahan et al. 2019, Mayfield et al. 2019). In cases where the CCN was available and questions were skipped, a deterministic record linkage protocol was designed and implemented in data processing to populate the resulting valid skips in the survey data with direct substitution of administrative data to create a blended data product as illustrated in Figure 1. The sections that follow provide further detail on the data sources and record linkage procedure.
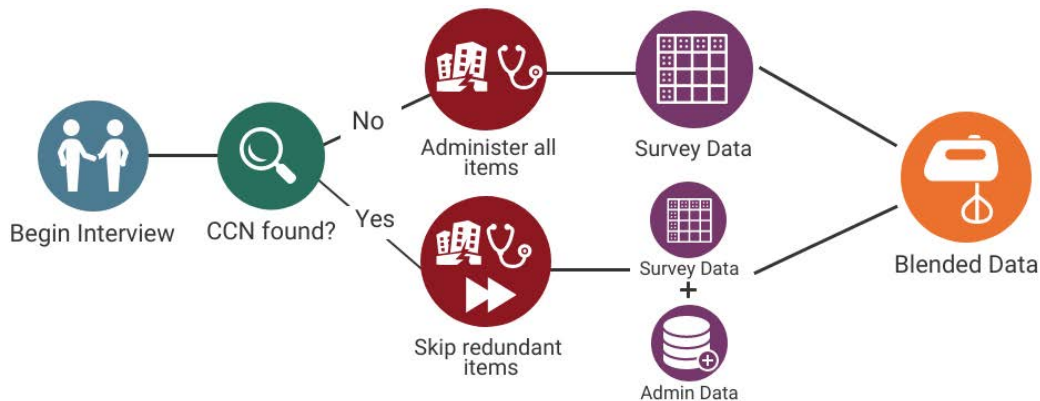

Figure 1: Redesigned MCBS Facility instrument flow and data processing protocol.

2.   Data Sources

Data processing incorporates round-based MCBS survey data from the FQ and HS questionnaire sections, as well as CASPER and MDS administrative data.

2.1 MCBS Survey Data

The data set storing information collected in the Facility Questionnaire (FQ) section of the MCBS Facility instrument consists of one record per beneficiary and the CCN (if collected) and other survey items pertaining to facility characteristics. The data set storing information collected in the Health Status (HS) contains one row per health assessment per beneficiary. The HS data set stores the CCN (if collected), the dates of any MDS forms reported by the facility respondent, and the type of assessment conducted. Importantly, this form metadata is collected even for cases reporting a CCN, so that it can later be used in linking to the MDS administrative data.

The survey data undergo the administrative matching process after each round of data collection, after cleaning and editing procedures have been executed but before the data have annualized for release and weights are calculated.

2.2 CMS Administrative Data

Table 1. CMS Administrative Data Sources

Dataset	Full Name	Description	Unit of observation
CASPER[1]	Certification and Survey Provider Enhanced Reporting	Required reporting of facility characteristics, e.g. bed counts, staffing, services provided.	1 record per CMS-certified nursing facility
MDS[2]	Long-Term Care Minimum Data Set	Federally-mandated health assessment of residents living in Medicare- and Medicaid-certified nursing homes.	1 record per Comprehensive or Quarterly MDS assessment

2.2.1 CASPER
CASPER supports the certification and regulatory function of CMS. Agencies perform regular surveys of facilities to determine whether the facility meets the requirements for participation in the Medicare and/or Medicaid programs. Certification survey data for every nursing home in the United States qualified to provide services under Medicare, Medicaid, or both are included in CASPER. As facilities are regularly resurveyed, new entries are added.

NORC obtains CASPER data from a vendor on an annual basis in the form of a cumulative file including records for facilities with both active and terminated CMS certifications. The CASPER database comprises seven unique data sets, two of which are used in MCBS data collection and data processing.

A provider-level file (CASPER Part 8) containing attributes such as the facility name, address, ownership, and accreditation is used to create the CCN questionnaire lookup tool used during data collection. Facilities classified as active providers located in the 50 states are eligible for inclusion in the lookup.

A certification-level file (CASPER Part 2) contains information such as bed counts, resident censuses, and services provided by the facility. The data in this file are used later to populate skipped questionnaire variables during data processing. Only the record corresponding to the most recent survey certification date for each facility is used.

2.2.1 MDS
The MDS is a federally-mandated health assessment of residents living in Medicare- or Medicaid-certified nursing homes. The purpose of the MDS is to assess and identify residents' health care problems, document individualized care plans, collect data for Medicare and Medicaid reimbursement systems, and monitor the quality of nursing home care. As such, the forms contain questions on numerous health-related topics, such as hearing, speech, and vision; cognitive patterns; mood; behavior; functional status; active diagnoses; health conditions, and medications.

---

[1] Cowles Research Group. (n.d.). CASPER & OSCAR. https://www.longtermcareinfo.com/data/casper-and-oscar.php
[2] ResDAC. (n.d.). Long Term Care Minimum Data Set (MDS) 3.0. https://resdac.org/cms-data/files/mds-30

There are two primary types of MDS assessments that are applicable to the MCBS – the Comprehensive assessment and the Quarterly assessment. Facility staff complete a Comprehensive MDS for residents upon their admission to a nursing home, then annually thereafter. The Comprehensive assessment is also completed if a resident experiences a significant change in health status. In addition, a subset of items from the MDS is completed for each resident on a quarterly basis to monitor changes in the resident's health status between comprehensive assessments (i.e. the Quarterly MDS).

MDS data are retrieved via the CMS Chronic Conditions Data Warehouse (CCW). MDS assessment data are available on a rolling basis approximately two months after data entry by facilities. MDS data are obtained at least two months after the conclusion of each round of data collection to help ensure alignment between the date ranges present in the MCBS survey data and the MDS administrative data.

The MDS data contain one record per MDS assessment conducted. Prior to data processing, the MDS data are restricted to Nursing Home Comprehensive and Nursing Home Quarterly assessments.

3. Record Linkage Protocol

3.1 FQ-CASPER Match
For beneficiaries living in a facility where a CCN was reported, the data set storing items collected in the FQ section can be linked directly to CASPER data using the CCN. The CCN reported in the Facility Instrument is matched to the most recent corresponding record on the CASPER Part 2 file with the same CCN.

The CCN questionnaire lookup tool is populated with data from the CASPER Part 2 file obtained at the same time as the CASPER Part 8 file, which ensures that matching administrative data will be available during data processing.
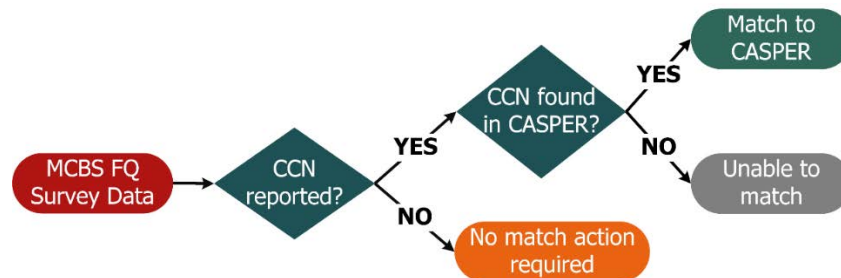

Figure 2. FQ-CASPER Match Diagram

Most variables are populated via direct substitution using a single CASPER variable that corresponds directly to the construct measured in the survey item, such as the number of Medicare-certified beds in a facility. Others are derived from multiple CASPER items. For example, a facility is classified as providing nursing or medical care if more than one resident receives any of the following types of care: a catheter, radiation therapy, chemotherapy, dialysis, intravenous therapy, respiratory treatment, tracheostomy care, ostomy care, suctioning, injections, or tube feedings.

3.2 HS-MDS Match
For beneficiaries living in a facility for which the facility respondent reported a CCN, the facility respondent is asked during the interview to identify the date and type

(Comprehensive or Quarterly) of any MDS assessments conducted for the sampled beneficiary on or around the survey reference date provided in the HS section. During data processing, the analytic file storing data collected in the HS section is linked to MDS data via a match protocol that identifies the "best" administrative data record, that is, the record most likely to be the MDS assessment reported by the facility respondent during the interview.

Prior to matching, the MDS data source is first restricted to records matching the case ID number and reported facility CCN for each beneficiary. If the case ID number and CCN combination cannot be found in the administrative data, the data are unable to be linked.
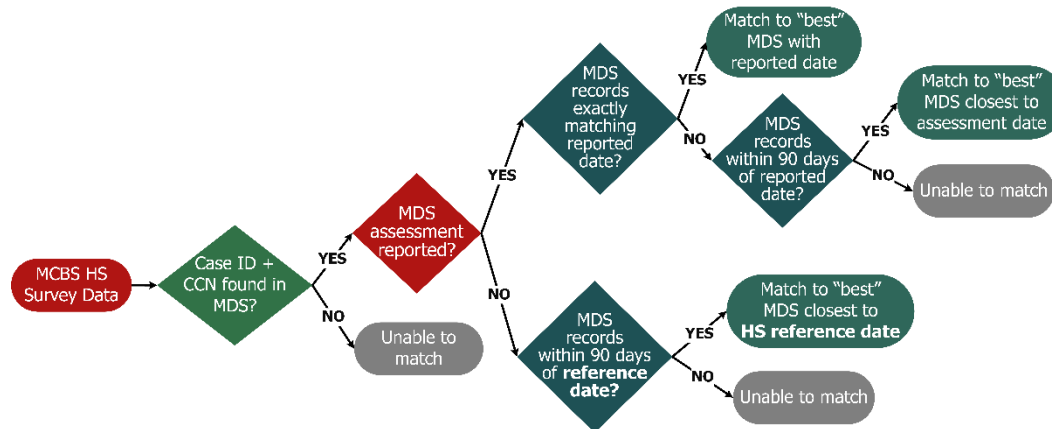

Figure 3. HS-MDS Match Diagram

The match algorithm prioritizes date matches, searching for MDS administrative data with the same target date as the survey-reported MDS assessment date. When there is no target date matching to the survey-reported assessment date, MDS administrative records with a target date within ±90 days of the assessment date are considered for matching.

If no MDS assessment is reported by the facility respondent in the interview, MDS administrative records with a target date within ±90 days of the HS survey reference date are considered for matching since a quarterly assessment is required to be conducted every 90 days. For most beneficiaries, the HS survey reference date is September 1 of the current year in the Fall round interview.

When multiple MDS administrative records corresponding to the same date are present, the "best" MDS record is chosen by comparing the survey-reported MDS assessment type (Comprehensive or Quarterly) to the MDS administrative assessment type. When multiple MDS administrative records within the ±90 date range are present, the "best" MDS record is chosen by first looking for the target date closest to the assessment or reference date, and then comparing the survey-reported MDS assessment type (Comprehensive or Quarterly) to the MDS administrative assessment type.

After linking the MDS record to the survey data, variables are populated via direct substitution using individual MDS items that correspond directly to the construct measured in the survey item. This is aided by the fact that the MCBS Facility instrument has been designed to mirror the MDS and references to MDS item numbers are included in the instrument itself. As such, the MDS data used are mostly comparable to the survey-

reported data, but values of dash, "-", that are used to signify missing data in the administrative source are converted to missing values consistent with the survey data.

4. Data and Methods

The record linkage protocol to integrate administrative data was implemented in MCBS data processing beginning in Fall 2019. Quality control measures designed to ensure the match process was executed accurately were performed each round. After completing two Fall rounds of data collection and data processing with the redesigned Facility instrument, the present study was initiated to explore the quality of the resulting blended data in greater depth.

Two analytic data sets were constructed. The first pools two years of FQ survey-only data (including responses abstracted by interviewers or staff) collected in Fall 2017 and Fall 2018 (n=2,190) and two years of FQ blended data created under the redesigned process in Fall 2019 and Fall 2020 (n=2,057). The same procedure was used to construct an HS analytic dataset containing pooled survey-only data from 2017 and 2018 (n=2,348) and blended data from 2019 and 2020 (n=2,163).

In the FQ data, there were 5 categorical variables and 4 continuous variables eligible for comparison between the survey-only and blended data, and 109 categorical variables and 3 continuous variables in the HS data. Some questionnaire variables that existed in these sections prior to 2019 were removed, changed, or could not be replaced with administrative data and, thus, are excluded from this comparison.

Analyses were performed to address three research questions covering different domains of data quality. The record linkage and all subsequent analyses were conducted in SAS Studio.

RQ1 (Match Quality). How complete and accurate was the record linkage for cases with a CMS Certification Number (CCN)? Using blended data only:
- Percent of FQ and HS records with a CCN successfully linked to administrative data
- Frequency of HS record match result (e.g. exact match, approximate match, no match found)
- Calculated date proximity between HS survey-reported and matched administrative MDS record date
- Percent agreement between HS survey-reported and administrative MDS form type (i.e. Comprehensive, Quarterly)

RQ2 (Comparability). How comparable are the survey-reported and blended statistics?
- Chi-square tests of independence for categorical variables
- T-tests for means of continuous variables
- F-test of equality of variances for continuous variables
- Descriptive statistics comparing minimum and maximum values of continuous variables

RQ3 (Item Non-Response). Does the record linkage process result in changes to item non-response or missingness?
- T-test for proportion of item non-response (missing, don't know, refused)

5. Results

5.1 Match Quality

The record linkage protocols produced a nearly perfect match rate. The facility questionnaire (FQ) records with a CCN reported were matched to CASPER data in 100% of cases. This rate was expected as the CCN questionnaire lookup tool is populated using the same CASPER administrative data file used in the match, ensuring that matching data will be available during data processing except in very rare scenarios.

Additionally, 96% of health status (HS) records with a CCN reported were successfully linked to an administrative MDS record, with the vast majority matched to a survey-reported MDS assessment rather than the survey reference date. For 90% of records linked using an MDS assessment date (77% of total records with a CCN), there was an exact match between the date of the MDS assessment reported by the facility staff member and the date of the linked MDS assessment in the administrative data. Table 2 contains the date proximity results for all HS records with a CCN.

Table 2: Date Proximity between Survey-Reported and Administrative MDS Assessment Date among Records with CCN

Match result	Percent	Frequency
Not matched	**4%**	**42**
Matched by survey-reported MDS assessment date	**86%**	**961**
Exact match	90%	865
1-7 days (1 wk)	4%	43
8-14 days (2 wks)	2%	17
15-30 days (1 mo)	2%	20
31-90 days (3 mos)	2%	16
Matched by survey reference date	**11%**	**117**
0-7 days (1 wk)	30%	35
8-14 days (2 wks)	20%	23
15-30 days (1 mo)	24%	28
31-90 days (3 mos)	26%	31
Total	**100%**	**1,120**

Among records successfully linked to MDS administrative data (by assessment date or reference date), 91% had agreement between the MDS assessment type in the administrative data and the assessment type reported during the interview (Table 3).

Table 3: Survey-Reported and Administrative MDS Assessment Type

Survey-reported MDS Type	Administrative MDS Type	Percent	Frequency
Comprehensive	**Comprehensive**	**65%**	**620**
Quarterly	**Quarterly**	**26%**	**252**
Comprehensive	Quarterly	3%	29
N/A	Comprehensive	2%	22
N/A	Quarterly	3%	24
Quarterly	Comprehensive	1%	14
Total		**100%**	**961**

Taking both date proximity and assessment type together, a total of n=810 (77%) of HS records with a CCN were matched exactly on assessment date and type. This suggests a high level of accuracy and completeness in both the survey-reported and administrative data. The majority of the time during the interview, facility staff and interviewers can

identify a specific MDS assessment conducted in the facility that can later be retrieved from administrative data during data processing.

5.2 Data Comparability

Previous feasibility research directly linked administrative data to existing MCBS survey data to compare parallel data points as reported in MDS/CASPER vs. the MCBS. This prior work found high levels of agreement between individual values in both datasets; approximately 80% of records had up to five discrepant data points, and fewer than 10% had greater than 10 discrepant values (Mayfield et al. 2019). The present study instead focused on the comparability of the blended data in its entirety to prior years' survey-only data to assess the impact of the match protocol, including any failures to link data, on the resulting estimates.

A minority of categorical variables had significantly different distributions in the blended data, using chi-square tests of independence and a significance level of $\alpha = .05$.

Among the five comparable categorical variables in the FQ data, one was significantly different in the blended data compared to the survey-only data. Among the 109 comparable categorical variables in the HS data, 37 (34%) were significantly different in the blended data compared to the survey-only data. Many of these variables (n=19) were also significant at $\alpha = .01$. The Appendix contains all variables and chi-square test results.

As for continuous variables, no statistically significant differences were observed between the means of the survey-reported and blended data for the four variables in FQ and three variables in HS. However, some differences in dispersion were observed in both datasets based on the F-test of equality of variances (Table 4).

Two FQ items, Medicaid-only beds (F=1.66, p < .001) and Medicare-only beds (F=1.78, p < .001), had significantly lower variance in the blended data than in the survey-reported data. Since the vast majority of beds are typically certified for both Medicare and Medicaid, we hypothesize that the number of beds only certified by Medicare or Medicaid may be less commonly known or misreported during the interview, while the administrative data has greater precision and accuracy. In contrast, two HS items, height (F=1.27, p < .001) and weight (F=1.14, p = .003) had a significantly higher variance in the blended data than in the survey-reported data. This is due in part to clearly erroneous values in the administrative MDS data, such as a height of 0 inches or a weight of 16 pounds; when these values were set to missing, height was no longer significant (F = 1.01, p = 0.89) though the significance of weight was unchanged (F=1.15, p=.001).

Table 4. Significant Differences in Variance for Continuous Variables

Table	Variable	Description	Survey Data Standard Deviation	Blended Data Standard Deviation
FQ	CAIDBEDS	Count of beds only certified for Medicaid	17.18	13.31
FQ	CAREBED	Count of beds only certified for Medicare	22.23	16.68
HS	HEIGHT	Beneficiary height (inches)	4.30	4.84
HS	WEIGHT	Beneficiary weight (lbs)	47.93	51.17

5.3 Item Nonresponse

Levels of missingness were assessed for all comparable variables. Item nonresponse was the same or lower in the blended data for facility characteristics (FQ). Of 10 variables in

FQ, item missingness was significantly different between the blended and survey-only data for three items. Overall missingness was low for most items in the survey-only data (<1%), but all variables were almost fully populated in the blended data (Figure 4). This suggests improved data quality as a consequence of incorporating administrative data, when compared to facility records used by the staff or field interviewer.
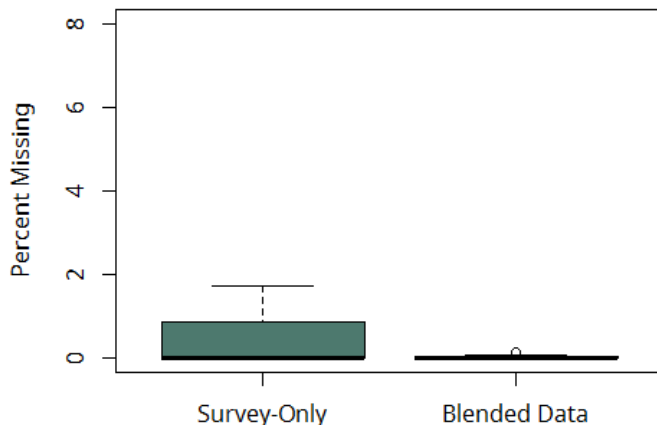

Figure 4: FQ Item Nonresponse

In contrast, item nonresponse was consistently higher in the health status (HS) blended data than in the survey-reported data. Of 114 variables in HS, item missingness was significantly different between the blended and survey-only data for 110 items (Figure 5). This is likely related to the 4% of HS records that could not be linked to administrative data; for these cases, the valid skips remained missing after data processing. Additionally, some items are less frequently populated in the MDS administrative data. Although the blended data for HS contains more missingness, we consider these levels (between 1% and 3% for most variables) to be within an acceptable threshold.
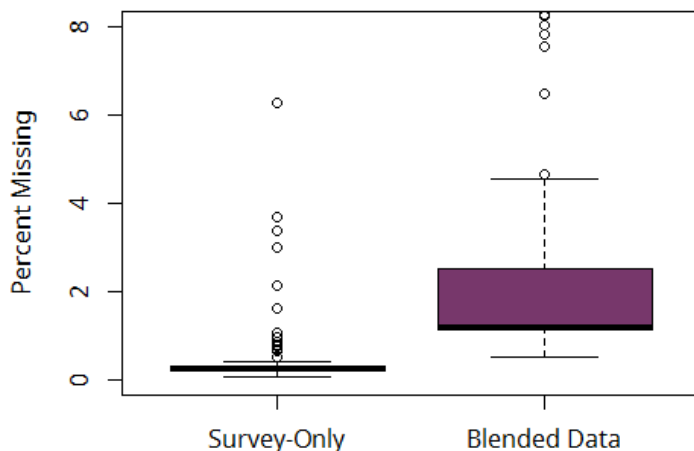

Figure 5: HS Item Nonresponse

6. Discussion

Overall, this record linkage process results in a near-complete match rate with largely comparable blended data while offering significant operational gains and a large reduction in respondent burden. The median MCBS Facility interview length decreased

in 2019, which could largely be attributed to the redesigned Facility instrument. Anecdotal feedback from field interviewers was overwhelmingly positive, emphasizing easier interview administration and the use of the CCN as a tool for gaining respondent cooperation. Additionally, once the linkage mechanism was established between the MCBS and the CASPER and MDS datasets, additional items from those same data sources could be integrated without increasing burden for cases with a CCN.

The deterministic record linkage protocol was highly successful for both the FQ and the more complex HS data, indicating the CMS administrative data is well aligned to both the universe and in time. Several future improvements could be explored, including changing the criteria for prioritizing assessments and follow-up investigation for records that could not be matched. Additionally, most variables were populated by direct substitution without adhering to questionnaire logic, and future work could examine whether administrative data follows the same skip patterns as the current instrument.

The results from the comparability and item missingness analyses suggest that data quality remains largely the same, particularly when weighed against the operational benefits of the process. However, a substantial number of variables, particularly in the HS data, have significant changes in their distributions that warrant further exploration. The U.S. Census Bureau's (2018) guiding principles for assessing uses of administrative data include whether differences might carry over to other variables indirectly through editing and imputation. Since our analyses were conducted on raw data, additional research is needed to determine the impact of these differences in the final data products, especially for derived variables (e.g. presence of chronic conditions) constructed from multiple survey items. Increased item nonresponse and unlikely or unusual values in the data should also be further analyzed to determine whether they create systematic bias.

The analysis has a few limitations. Trends in facility characteristics and beneficiary health status may have changed in the time since the 2017 data was collected, which would impact any comparisons, and demographics were not controlled for in the analyses. Additionally, the lack of experimental design and knowledge of the "true" errors in the data limits our ability to make definite conclusions about data quality. Interviewer-related error can take a number of forms in survey response, including data entry errors and interviewer behavior (Lavrakas, 2008). In the MCBS Facility interview, interviewers and facility staff abstract from facility medical records, which can be complicated by navigating complex forms and making subjective determinations of how to match available information to the requested constructs (Zozus et al. 2019). Although the CMS data sources are generally assumed to be of high quality due to their use in certification, data entry errors or delays can occur depending on the size of and resources in the facility, and in rare instances there have been reports of data falsification (Silver-Greenberg and Gebeloff, 2021). Continued research is needed to better understand the quality and precision of survey-reported, abstracted, and administrative data.

For data systems with access to reliable and high-quality administrative data, including the federal statistical ecosystem, this process offers a valuable methodology to reduce burden while maintaining data quality. The availability of unique identifiers and the collection of key variables for matching within the instrument allows for successful implementation of a relatively simple deterministic record linkage protocol with a very high match rate. The volume of available administrative data is growing over time, with increased attention its accuracy and timeliness. These continued advancements should lead to regular reviews of opportunities for leveraging administrative data in surveys.

Disclaimer

The opinions and views expressed in this work are those of the authors. No official endorsement by the Department of Health and Human Services or the Centers for Medicare & Medicaid Services is intended or should be inferred.

References

Auerbach, J., Brummet, Q., Czajka, J., Hough, G. C., Hunsinger, E., & Salvo, J. (2019). Will administrative data save government surveys?. *Significance*, *16*(5), 35-39.

Carnahan, R., Mayfield, A., LeClere, F., & Rosner, S. (2019, April 16). *Designing a survey tool to determine when administrative data can substitute for survey-collected data* [Conference presentation]. FedCASIC 2019, Washington, DC, United States. https://www.census.gov/fedcasic/fc2019/ppt/2CCarnahan.pdf

Davern, M. E., Meyer, B. D., & Mittag, N. K. (2019). Creating improved survey data products using linked administrative-survey data. *Journal of Survey Statistics and Methodology*, *7*(3), 440-463.

Gordon, J., Eckman, S., Einaudi, P., Sanders, H., & Yamaner, M. (2018). Using administrative records to increase quality and reduce burden in the Survey of Graduate Students and Postdoctorates in Science and Engineering. *Statistical Journal of the IAOS*, *34*(4), 529-537.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly*, *74*(5), 849-879.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., ... & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, *79*(4), 839-880.

Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, *32*(2), 293-312.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods* (Vols. 1-0). Thousand Oaks, CA: Sage Publications, Inc. doi: 10.4135/9781412963947

Mayfield, A., Carnahan, R., and F. LeClere. (2019). Integrating administrative data with survey-collected data to reduce burden in establishment data collection. *JSM Proceedings*, Administrative Data Section, 781-794.

Seeskin, Z. H., LeClere, F., Ahn, J., & Williams, J. (2018). Uses of alternative data sources for public health statistics and policymaking: challenges and opportunities. *JSM Proceedings*, Government Statistics Section, 1822-1861.

Silver-Greenberg, J., & Gebeloff, R. (2021, March 13). *Maggots, rape and yet five stars: How U.S. ratings of nursing homes mislead the public*. The New York Times. https://www.nytimes.com/2021/03/13/business/nursing-homes-ratings-medicare-covid.html.

U.S. Census Bureau. (2015, June). *Agility in Action: A Snapshot of Enhancements to the American Community Survey*. https://www.census.gov/content/dam/Census/programs-surveys/acs/operations-and-administration/2015-16-survey-enhancements/Agility_in_Action.pdf.

U.S. Census Bureau. (2018, November). *Realizing the Promise of Administrative Data for Enhancing the American Community Survey*. https://www.census.gov/content/dam/Census/programs-surveys/acs/operations-and-administration/administrative-records-in-the-american-community-survey.pdf.

Zozus, M. N., Young, L. W., Simon, A. E., Garza, M., Lawrence, L., Ounpraseuth, S. T., ... & Devlin, L. A. (2019). Training as an intervention to decrease medical record abstraction errors multicenter studies. *Studies in health technology and informatics*, *257*, 526.

Appendix: Significant Chi-Square Test Results

Table	Variable	Description	χ^2	P-value
FQ	NORMCARE	Facility provides nursing or medical care		0.002
HS	HCHEAID	SP had hearing aid	7.99	0.005
HS	MENTCON	Interview for mental status conducted	7.90	0.005
HS	PFTRNSFR	SP's level of self-performance in transferring	20.05	0.003
HS	PFLOCOMO	SP's level of self-performance in locomotion	33.07	<.001
HS	CTBOWELC	SP's level of bowel control	11.38	0.023
HS	CANCER	SP had cancer	9.98	0.002
HS	AFIBDYS	SP had atrial fibrillation or other dysrhythmia	6.85	0.009
HS	CORARTDS	SP had coronary artery disease	22.94	<.001
HS	GERDULUC	SP had GERD or ulcer	13.25	<.001
HS	RENLESRD	SP had renal insufficiency, failure, or ESRD	26.44	<.001
HS	NUROBLAD	SP had neurogenic bladder	5.07	0.024
HS	HYPERKAL	SP had hyperkalemia	4.45	0.035
HS	ARTHRIT	SP had arthritis	19.88	<.001
HS	APHASIA	SP had aphasia	5.40	0.020
HS	CERPALSY	SP had cerebral palsy	15.10	<.001
HS	CVATIAST	SP had CVA, TIA, or stroke	4.07	0.044
HS	DEMENT	SP had dementia	6.08	0.014
HS	HEMIPLPA	SP had hemiplegia/hemiparesis	14.59	<.001
HS	SEIZEPIL	SP had seizure disorder or epilepsy	7.63	0.006
HS	MALNUTRI	SP had malnutrition or was at risk for it	29.36	<.001
HS	MANICDEP	SP had manic depression	6.53	0.011
HS	PSYCOTIC	SP had psychotic disorder other than schizophrenia	3.99	0.046
HS	PTSD	SP had post-traumatic stress disorder	4.96	0.026
HS	RESPFAIL	SP had respiratory failure	7.28	0.007
HS	SOLOSS	SP experienced a loss from mouth when eating/drinking	7.44	0.024
HS	SOHOLD	SP experienced problem holding food in mouth	6.86	0.032
HS	SOPAIN	SP experienced pain or difficulty with swallowing	6.15	0.046
HS	SODENT	SP experienced problem with dentures	7.89	0.019
HS	SOTEETH	SP had no natural teeth or fragments	31.44	<.001
HS	SOTISSUE	SP had abnormal mouth tissue (ulcer/mass/lesion)	7.70	0.021
HS	SOCAVITY	SP had obvious likely cavity or broken teeth	17.13	<0.001
HS	SOGUMS	SP had inflamed gums or loose teeth	7.73	0.021
HS	SOCHEW	SP had mouth pain or problem with chewing	7.99	0.018
HS	SONOEXAM	Unable to examine SP for swallowing/oral problems	9.76	0.008
HS	INFMDRO	SP had multidrug-resistant organism (MDRO)	3.92	0.048
HS	INFSEPT	SP infected by septicemia	4.29	0.038
HS	INFURNRY	Urinary tract infection occurred	10.30	0.001