

Chronic Conditions Warehouse

Your source for national CMS Medicare and Medicaid research data



Chronic Conditions Warehouse Virtual Research Data Center

Medicare Bayesian Improved Surname Geocoding (MBISG) File User Guide

MARCH 2026 | VERSION 1.3

Revision Log

Date	Changed by	Revisions	Version
March 2026	D. Happe	Updated ResDAC email to DataRequests@cms.hhs.gov	1.3
April 2025	D. Happe	Edits made to comply with Executive Order 14168	1.2
January 2025	K. Schneider	Updated document to include 2024 data file	1.1
March 2024	K. Schneider	Created initial document	1.0

Table of Contents

1.0 Overview	1
2.0 Background	2
3.0 CCW MBISG File	3
3.1 Race and Ethnicity Probabilities	3
3.2 Race and Ethnicity Classification.....	3
3.3 Predicted Spanish Preference Category.....	4
3.4 Intended Use for MBISG 2.1.2	4
3.5 Guidance for Performing Estimation Using MBISG R/E Probabilities:	5
3.6 Guidance for Performing Estimation Using the MBISG Classification:	6
4.0 Linking with Other CCW Data Files	9
4.1 Medicare Part A, B, C, and D Enrollment Segment.....	9
4.2 Medicare Part A and B Claims	10
5.0 Receiving CCW Data	11
5.1 Within the CCW Virtual Research Data Center (VRDC).....	11
5.2 Physical Shipment of Data	11
6.0 Where to Get Assistance	13
Appendix A — List of Acronyms	14

List of Tables

Table 1. MBISG 2.1.2 file variables.....	3
Table 2. FFS claims for diabetes weighted for Black race and ethnicity.....	8
Table 3. Format and naming convention for the CCW files.....	12
Table 4. MBISG SDA contents	12
Table 5. CCW resources accompanying data files.....	12

List of Data Tips

Data Tip 1. Estimating the proportion of beneficiaries in an MBISG subgroup with diabetes, according to fee-for-service data.....	6
--	---

1.0 Overview

Medicare is the primary health insurance program for people aged 65 or older, people under age 65 with disabilities, and people of all ages with end-stage renal disease (ESRD). The Centers for Medicare & Medicaid Services' (CMS) Office of Minority Health (OMH) modified an existing method for indirectly estimating race and ethnicity from surname and residential information (Bayesian Improved Surname and Geocoding [BISG]) to augment CMS's Social Security Administration (SSA)-based administrative measure of race; this resulted in the Medicare Bayesian Improved Surname Geocoding (MBISG) algorithm.

CMS uses the Chronic Conditions Warehouse (CCW) to develop and manage CMS research data resources. The CCW has complete (100%) Medicare enrollment, fee-for-service (FFS) claims data, and the MBISG 2.1.2 data file obtained directly from CMS. From this source data, the CCW team has prepared a data file to disseminate to researchers and certain government agencies that CMS has approved under a Data Use Agreement (DUA) to obtain MBISG data for research purposes. The CCW MBISG data file contains identifiable information. It is subject to the Privacy Act and other federal government rules and regulations (reference the Research Data Assistance Center [ResDAC] website for details on requesting Medicare data <http://www.resdac.org/>).

This guide provides researchers with information to clarify their work with the CCW MBISG data file. [Appendix A](#) lists abbreviations used in this document.

2.0 Background

CMS developed the MBISG algorithm to augment the SSA-based measure of race. The BISG method combines U.S. Census Bureau data on race and ethnicity distributions by surname and census block group to produce a set of probabilities of falling into each of six racial and ethnic groups: American Indian or Alaska Native (AI/AN), Asian American and Native Hawaiian or Other Pacific Islander (AA and NHPI), Black, Hispanic, Multiracial, and White. The Medicare-specific adaptation — MBISG — combines the BISG race and ethnicity probabilities with CMS's race and ethnicity administrative data to produce more accurate, indirect estimates of the race and ethnicity of Medicare beneficiaries. The CMS administrative data for this purpose includes the beneficiary's race code, first name, demographics, and coverage characteristics.

Research on ways to improve MBISG racial and ethnic estimates is ongoing. Periodically, CMS releases a new version of MBISG estimates, generally restricted to people currently enrolled in Medicare. In years when CMS does not release a new version, CMS releases an updated file that adds estimates for people new to Medicare. In 2023, CMS released MBISG 2.1.2.

Researchers can find more detail on the MBISG methodology in this [Health Services Research](#) article.¹

¹ Haas A, Elliott MN, Dembosky JW, Adams JL, Wilson-Frederick SM, Mallett JS, Gaillot S, Haffer SC, Haviland AM. Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. *Health Services Research*. 2019 Feb;54(1):13-23. <https://doi.org/10.1111/1475-6773.13099>. (Accessed 01/08/2025)

3.0 CCW MBISG File

CMS uses the CCW to develop and manage CMS research data resources. The CCW team receives the MBISG data file from CMS and disseminates it. The CCW team joins the source MBISG data with the BENE_ID variable, allowing linkage to other CCW data products (e.g., Medicare enrollment and claims).

Throughout this chapter, the CCW team writes SAS variable names in all capital letters.

Periodically, CMS releases a new version of MBISG estimates; the current version is MBISG 2.1.2. This cross-sectional analytic file contains a snapshot of all Medicare beneficiaries enrolled on March 1 of the year. The first year of MBISG enrollment data CMS disseminated was 2023 (i.e., it identified beneficiaries alive and enrolled on March 1, 2023). The MBISG file contains a total of 11 variables (reference [Table 1](#)), including:

- CCW beneficiary ID
- MBISG 2.1.2 race and ethnicity (R/E) probabilities (six variables)
- MBISG 2.1.2 race and ethnicity classification
- Predicted Spanish preference category
- MBISG version number, and
- Reference date for data file

Table 1. MBISG 2.1.2 file variables

Long SAS name	Label
BENE_ID	Encrypted CCW beneficiary ID
RE_PRBLTY_AIAN	R/E probability — American Indian or Alaska Native
RE_PRBLTY_AANHPI	R/E probability — Asian American and Native Hawaiian or Other Pacific Islander
RE_PRBLTY_BLACK	R/E probability — Black
RE_PRBLTY_HSPNC	R/E probability — Hispanic
RE_PRBLTY_MULTIRACIAL	R/E probability — Multiracial
RE_PRBLTY_WHT	R/E probability — White
RE_CLSFCTN	R/E classification with highest probability
SPNSH_PREFNC_CTGRY	Predicted Spanish preference category
MBISG_VRSN	Version of the MBISG algorithm
RFRNC_DT	Data file reference date

3.1 Race and Ethnicity Probabilities

CMS assigns each Medicare beneficiary in this file a set of probabilities of belonging to each of the six racial and ethnic groups. Researchers can reference these probabilities by the following variables: RE_PRBLTY_AIAN, RE_PRBLTY_AANHPI, RE_PRBLTY_BLACK, RE_PRBLTY_HSPNC, RE_PRBLTY_MULTIRACIAL, and RE_PRBLTY_WHT. Probabilities sum to 1 for each beneficiary, represented by a record in the file.

3.2 Race and Ethnicity Classification

To construct the classification variable (called RE_CLSFCTN), CMS initially classifies each person to the racial and ethnic group with the highest MBISG 2.1.2 probability across their set of six probabilities. CMS classifies very few people into the multiracial group using this rule, and classification into this group is not strongly predictive of self-reported race and ethnicity. For these reasons, the multiracial classification will not be useful in most applications. Therefore, CMS

classifies people for whom “Multiracial” is the highest probability into the group indicated by their second highest MBISG probability.

3.3 Predicted Spanish Preference Category

CMS estimates the probability that each person prefers Spanish-language survey material using both their Hispanic MBISG 2.1.2 indirect estimate (RE_PRBLTY_HSPNC) and the proportion of residents in their county who are Spanish-speaking with limited English proficiency (according to the 2017–2021 American Community Survey [ACS] for the 2023 MBISG file, and the 2018–2022 ACS for the 2024 MBISG file). CMS bases the predictive model on the results of a randomized experiment of survey methods involving Medicare beneficiaries.²

CMS categorizes predictions by the SPNSH_PREFNC_CTGRY variable into four groups ranging from high probability (category 1) to very low probability (category 4) of preferring Spanish language material. CMS categorizes residents of Puerto Rico into the highest-probability group (category 1). Predictions are unavailable for residents of other U.S. territories or people whose geographic data was missing in the input dataset used for calculations.

3.4 Intended Use for MBISG 2.1.2

MBISG 2.1.2 is a tool meant for describing the racial and ethnic composition of the Medicare population or subsets (e.g., members of specific Medicare Advantage contracts) and comparing health and health care by racial and ethnic groups. MBISG 2.1.2 is unsuitable for making statements about the race and ethnicity of individual Medicare beneficiaries.

MBISG 2.1.2 imputations strongly predict self-reported race and ethnicity for AA and NHPI, Black, Hispanic, and White race and ethnicity. Researchers commonly measure predictive accuracy using the C-statistic, also called the concordance statistic or area under the curve, a common metric for the performance of classification models. The C-statistic ranges from 0.5 (no predictiveness) to 1.0 (perfect predictiveness), with 0.7, 0.8, and 0.9 corresponding to adequate, excellent, and outstanding prediction, respectively.³ C-statistics for MBISG 2.1.2 imputations of AA and NHPI, Black, Hispanic, and White race and ethnicity are 0.99, 0.99, 0.97, and 0.97, respectively. The C-statistic for AI/AN race and ethnicity is 0.84, which supports use with caution, ideally with effective sample sizes of at least 300. CMS does not currently recommend making inferences using the MBISG multiracial probability estimate (RE_PRBLTY_MULTIRACIAL).

MBISG 2.1.2 probabilities or classifications can provide insights about, for example, the racial and ethnic distribution of patients at a given hospital or members of a given plan. CMS calibrates the probabilities such that the sum of one of the racial and ethnic probabilities across plan members estimates the total number of people in the plan who are in that racial and ethnic group (this will generally not be a whole number). Similarly, the average of the probabilities for a given racial and ethnic group estimates the proportion of members who are part of that racial and ethnic group.

Researchers can use MBISG 2.1.2 probabilities or classifications to compare the quality of health care received by racial and ethnic groups. For example, researchers have used MBISG to compare Healthcare Effectiveness Data and

² Elliott MN, Klein DJ, Kallaur P, Brown JA, Hays RD, Orr NE, Zaslavsky AM, Beckett MK, Gaillot S, Edwards CA, Haviland AM. Using Predicted Spanish Preference to Target Bilingual Mailings in a Mail Survey with Telephone Follow-up. *Health Services Research* 2018; 54(1): 5-12. DOI: <https://doi.org/10.1111/1475-6773.13088> (Accessed 01/08/2025)

³ Hosmer DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*. 2013. Print.

Information Set (HEDIS) measures by racial and ethnic groups nationally and within plans.⁴ There are several ways to implement estimation of differences between groups.

3.5 Guidance for Performing Estimation Using MBISG R/E Probabilities:

- Researchers can implement the estimation of disparities via multiple imputations. In this approach, analysts create multiple datasets with each person coded into a single racial and ethnic group in each dataset, with the probability of assignment to each group reflecting that person's set of racial and ethnic probabilities from the variables in this file. The same person may appear in different groups in different datasets. Researchers do analyses (such as regression) separately on each replicate dataset and combine results according to Rubin's Rules; this approach makes weaker assumptions than using probabilities as regressors.⁵
- Researchers can use the MBISG racial probabilities as weights to estimate unadjusted means or proportions for racial and ethnic groups by using the probability of each racial and ethnic group in turn as a weight (reference the [Data Tip 1](#) later in the document). Researchers can use weighted regression for unadjusted or adjusted comparisons of groups. For weighted regression, researchers create a dataset in which every person has six records, one for each racial and ethnic group in this data file, with the record coded as belonging to that group and the corresponding MBISG probability attached to the record. Researchers then regress the outcome on race and ethnicity using the racial and ethnic probabilities as weights. Researchers approximate standard errors that account for the weighting by the probabilities and adjust for multiple observations. Within each person, researchers estimate using a sandwich estimator,⁶ with the person treated as the clustering factor; researchers can also compute bootstrap standard errors.
- It is possible to conduct regression using the MBISG racial and ethnic probabilities as predictors. To estimate differences between racial and ethnic groups, researchers can parameterize race and ethnicity with an omitted reference group (usually White when that is the largest group overall in the dataset). Researchers can use post-estimation procedures to compare any pair of groups. Researchers can set up models to compare each racial and ethnic group against all other groups or the overall mean of the outcome variable. Researchers can use these regressions with or without covariates, depending on whether absolute or adjusted differences are of interest. In terms of scale, researchers can interpret the coefficients on the probabilities as if they came from 1/0 coding of the same variables, with the same reference group, under conditions specified in McCaffrey and Elliott (2008).⁷ Researchers can use recycled predictions from these models to estimate means of outcome variables by race and ethnicity.^{8,9}

⁴ Haas A, Elliott MN, Dembosky JW, Adams JL, Wilson-Frederick SM, Mallett JS, Gaillot S, Haffer SC, Haviland AM. Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. *Health Services Research*. 2019 Feb;54(1):13–23. <https://doi.org/10.1111/1475-6773.13099> (Accessed 01/08/2025)

⁵ McCaffrey DF, Elliott MN. Power of tests for a dichotomous independent variable measured with error. *Health Services Research*. 2008;43(3):1085–1101. <https://doi.org/10.1111/j.1475-6773.2007.00810.x> (Accessed 01/08/2025)

⁶ Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 2015. 50(2), 317-372.

⁷ McCaffrey DF, Elliott MN. Power of tests for a dichotomous independent variable measured with error. *Health Services Research*. 2008;43(3):1085–1101. <https://doi.org/10.1111/j.1475-6773.2007.00810.x> (Accessed 01/08/2025)

⁸ Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res*. 2008 Oct;43(5 Pt 1):1722-36. doi: 10.1111/j.1475-6773.2008.00854.x. Epub 2008 May 12. PMID: 18479410; PMCID: PMC2653886.

⁹ Haas A, Adams JL, Haviland AM, Dembosky J, Morrison PA, Gaillot S, Fremont A, Gildner J, Tamayo L, Elliott MN. The contribution of first-name information to the accuracy of racial and ethnic imputations varies by sex, race, and ethnicity. *Medical Care* 2022;60(8):556–562 <https://doi.org/10.1097/MLR.0000000000001732> (Accessed 03/31/2025)

3.6 Guidance for Performing Estimation Using the MBISG Classification:

Using probabilities as weights¹⁰ or as regressors¹¹ approximates the multiple imputation approach.¹² Approaches explained by Dembosky, et. al.¹³ and especially Cameron and Miller¹⁴ require stronger assumptions than the multiple regression approach documented in the CMS paper.¹⁵ Although the better approximation of the multiple imputation approach may vary by application, and results are often similar, simulations suggest that the weighting approach may be more robust than the regressors approach to violations of its assumptions.

Because MBISG probabilities are strongly bimodal (98% of people have one probability > 0.70 and 90% have one probability > 0.90), the classification (RE_CLSFCTN) contains most but not all the information in the probabilities. Hence, it is possible to use the classification variable in analyses as if the researcher knows the race and ethnicity. This approach is simple to implement but somewhat understates standard errors and is not as accurate as multiple imputation. In addition, this approach underestimates the prevalence of lower prevalence groups if used to estimate the size of groups by race and ethnicity, for which sums of probabilities will be more accurate.

The CCW team presents the following example to demonstrate the appropriate use of information in the MBISG data file. The MBISG probability weights work as intended in the overall Medicare population or subsamples, such as only people in Medicare Advantage (MA) or only people with FFS coverage. Researchers can use the MBISG probability weights for smaller samples, such as estimating the racial and ethnic distribution of people in a particular MA plan.

For simplicity, the CCW team assumes investigators would identify a study population using Medicare enrollment data or other person-level data (such as MBSF_ABCD or MBSF_CHRONIC). However, researchers may identify populations using different files or criteria.

Data Tip 1. Estimating the proportion of beneficiaries in an MBISG subgroup with diabetes, according to fee-for-service data

The CCW makes it easy to study chronic diseases by incorporating variables for common chronic conditions into the Medicare Master Beneficiary Summary File (MBSF) 30 CCW Chronic Conditions file segment (called the MBSF_CHRONIC_YYYY). The MBSF Chronic Conditions segment uses FFS claims-based algorithms to indicate treatment for a condition appears to have occurred; therefore, researchers cannot determine whether providers treated the managed care enrollees for the condition(s) of interest. The data file includes a yearly variable that indicates whether the beneficiary met the CCW chronic condition during the look-back period. For this example, the CCW team identifies FFS Medicare beneficiaries in the MBSF_CHRONIC_YYYY file for 2022 who were receiving diabetes treatment.

¹⁰ Centers for Medicare & Medicaid Services. Office of Minority Health. Stratified Reporting. <https://www.cms.gov/About-CMS/Agency-Information/OMH/research-and-data/statistics-and-data/stratified-reporting> (Accessed 01/08/2025)

¹¹ Dembosky JW, Haviland AM, Haas A, Hambarsoomian K, Weech-Maldonado R, Wilson-Frederick SM, Gaillot S, Elliott MN. Indirect Estimation of Race/Ethnicity for Survey Respondents who do not Report Race/Ethnicity. *Medical Care*. 2019 May;57(5):e28–e33. <https://doi.org/10.1097/MLR.0000000000001011> (Accessed 01/08/2025)

¹² Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. *Journal of human resources*, 2015. 50(2), 317–372.

¹³ Dembosky, et. al., loc. cit.

¹⁴ Cameron, et. al., loc. cit.

¹⁵ CMS, loc. cit.

Using any of the six race and ethnicity probabilities in the MBISG, researchers can weigh the chronic conditions data to obtain Medicare FFS population estimates for the racial and ethnic group(s) who have diabetes. For this example, the CCW team uses the 2023 MBISG file and the Black racial group probability as a weight.

A SAS® coding example illustrates one method for accomplishing this task; of course, researchers may adapt this code to use any statistical software they prefer:

1. Begin by identifying FFS Medicare beneficiaries with diabetes. To identify FFS beneficiaries, the CCW team uses the MBSF_ABCD_2022. For simplicity in identifying beneficiaries with diabetes, they used the pre-coded diabetes condition variable in the MBSF_CHRONIC_2022.
2. Next, merge the diabetes information from the annual MBSF_CHRONIC file (the [CCW Technical Guidance: Calculating Medicare Population Statistics](#) document provides greater detail and examples for calculating population statistics).
3. Then joins the MBISG data (MBISG_MARCHSNAPSHOT_2023) to the diabetes population data by BENE_ID. Since the beneficiaries in these two datasets do not match 1:1, the CCW team suggests using a left join when merging these files – keeping only MBISG records if there is a corresponding MBSF or enrollment record.
4. Calculate a weighted frequency for Medicare FFS beneficiaries with diabetes for the racial group(s) of interest.

When the variable and filenames are in all capital letters, they come directly from CCW data files, whereas the lower-case variable and filenames are derived variables or temporary files.

```

/*identify FFS Medicare population with diabetes*/
data FFS_Diab_2022;
merge MBSF.MBSF_ABCD_2022 (keep=BENE_ID MDCR_ENTLMT_BUYIN_IND_01 -
MDCR_ENTLMT_BUYIN_IND_12 HMO_IND_01 - HMO_IND_12 BENE_death_dt BENE_RACE_CD
RTI_RACE_CD)
MBSF.MBSF_CHRONIC_2022 (keep=BENE_ID DIABETES);
  by BENE_ID;
      *note this input file is a merged MBSF A/B/C/D and MBSF_CC
Conditions determine # months of Part A, B, and no HMO coverage;
array MemberMos_AB (12)
MDCR_ENTLMT_BUYIN_IND_01 - MDCR_ENTLMT_BUYIN_IND_12;
array MemberMos_noHMO (12) HMO_IND_01 - HMO_IND_12;
array Member_FFSMos (12) Member_FFSMos01 - Member_FFSMos12;
do i = 1 to 12;
if MemberMos_AB(i) in ('3','C') and MemberMos_noHMO(i) in
('0','4')then Member_FFSMos(i)= 1;
else if MemberMos_AB(i) NOT in ('3','C') or MemberMos_noHMO(i) NOT in
('0','4')then Member_FFSMos(i)= 0;
Member_Mos=sum(of Member_FFSMos:);
end;
* determine who had 11- or 12-months FFS coverage or coverage until month
before death;
if (BENE_DEATH_DT=. and Member_Mos in (11,12)) or (BENE_DEATH_DT~=.
and month(BENE_DEATH_DT)<=Member_Mos+1 and Member_mos~=0) then FFS_Cov=1;
else FFS_Cov=0;

* numerator information for Diabetes – keep both 3s and 1s;
if FFS_Cov=0 then Diab=.;
else if FFS_Cov=1 and (DIABETES= 0 or DIABETES=2)then Diab=0;

```

```

else if FFS_Cov=1 and (DIABETES= 1 or DIABETES=3)then Diab=1;
label
FFS_Cov = '11 or 12 months FFS no HMO - except for those who died'
Member_Mos = 'Total Member months of A B and No HMO - per bene'
Diab = 'FFS claims for Diabetes';
run;

/* LEFT JOIN diabetes file to MBISG*/

proc sql;
create table diab_race
as select a.BENE_ID, a.diab, a.FFS_cov, b.*

from FFS_DIAB_2022 a left join MBISG_R.MBISG_MARCHSNAPSHOT_2023 b
on a.BENE_ID=b.BENE_ID;
quit;

proc freq data=diab_race;
weight RE_PRBLTY_BLACK;
table diab;
title 'weighted for BLACK';
run;

```

Results of this data exercise are:

Table 2. Beneficiaries with FFS claims for diabetes, weighted for Black race and ethnicity

Diab	Frequency*	Percent	Cumulative frequency	Cumulative percent
0	1153970	62.99	1153970	62.99
1	678137.1	37.01	1832107	100.00

* Frequency missing = 4610798.6345

These MBISG-weighted results show 36.08% of the black FFS Medicare population had diabetes in 2022. This code example sets non-FFS beneficiaries to missing; however, there are also missing observations in the results for beneficiaries in the 2022 MBSF who are not in the 2023 MBISG.

4.0 Linking with Other CCW Data Files

The beneficiaries in the MBISG 2.1.2 file are Medicare beneficiaries who appear in the CMS enrollment source data to have been alive and enrolled on March 1 of the reference year. For example, the 2023 MBISG file is the snapshot of all beneficiaries enrolled in March 2023.

CCW adds a unique CCW beneficiary identifier (the BENE_ID) in each data file delivered as part of the output package. The unique CCW beneficiary identifier provides a common link across all available data types, thus allowing data users to link the MBISG data to beneficiary and claims data in the CCW.

The unique CCW beneficiary identifier field is specific to the CCW and does not apply to any other identification system or data sources. CCW encrypts this identifier and all data files before delivering the data files to researchers.

4.1 Medicare Part A, B, C, and D Enrollment Segment

The CCW Medicare enrollment data file is the Master Beneficiary Summary File (MBSF) sourced from the CMS Common Medicare Environment (CME) database. The MBSF contains many enrollment and other person-level variables in file “segments.” These segments are separate components of the file researchers may request. The [data dictionaries](#) on the CCW website describe the variables contained in the MBSF.

The CCW team creates the MBSF for each calendar year. The MBSF contains demographic entitlement and enrollment data for beneficiaries who 1) CMS documents are alive for some of the reference year and 2) enrolled in the Medicare program during the file’s reference year. Reference year refers specifically to the calendar year accounted for in the MBSF. So, for example, the 2022 MBSF covers the year 2022 — that is the reference year.

This essential information for most study denominators appears in the Base A/B/C/D segment of the MBSF. For each of the MBSF file segments, there is one record for each BENE_ID. The additional segments of MBSF are 1) CCW Chronic Conditions, 2) CMS Other Chronic or Potentially Disabling Conditions (OTCC), 3) Cost and Use, and 4) National Death Index (NDI).¹⁶

Researchers may wish to obtain MBSF data fields for the beneficiaries in the MBISG data file. Remember, the MBISG identifies beneficiaries using a snapshot of all Medicare eligibles at a point in time. There is no 1:1 match between beneficiaries in the MBSF and the MBISG file. The first release of the MBISG version 2.1.2 file identifies all beneficiaries enrolled on March 1, 2023; the 2024 MBISG identifies all beneficiaries enrolled on March 1, 2024. The CCW team calls this variable in the MBISG file the data file reference date (RFRNC_DT). Some beneficiaries in the 2022 MBSF will not appear in the 2023 MBISG (for example, if they died between December 31, 2022, and March 1, 2023). Similarly, CMS does not include some beneficiaries who are in the 2023 MBISG file in the 2022 MBSF (for example, if beneficiaries newly enrolled in Medicare in 2023). Use the BENE_ID to link the MBSF and the MBISG.

¹⁶ Researchers may only use the NDI files within the CCW Virtual Research Data Center (VRDC).

4.2 Medicare Part A and B Claims

The CCW system includes Medicare institutional and non-institutional claims and Medicare Part D prescription drug events. CMS historically limited the Medicare claims found in the CCW to FFS Part A and B claims only. The [Data Dictionaries](#) tab on the CCW website describes the variables in the FFS claims files; researchers may also reference the [CCW Medicare Administrative Data User Guide](#) on the CCW website.

MA (Part C) encounter data RIFs are available to researchers starting with 2015. Medicare Advantage Organizations (MAOs) are private managed care plans, such as health maintenance organizations (HMOs), preferred provider organizations (PPOs), private fee-for-service plans (PFFS), and special needs plans (SNPs) that provide Medicare Part A and Part B services. MAOs submit data to CMS that the CCW team uses to create the RIFs. Reference the [Medicare Encounter records data dictionaries](#) and the [CCW Medicare Encounter Data User Guide](#) on the CCW website.

The CCW team adds key variables in the data files to help researchers join them together as appropriate (e.g., the unique CCW-assigned beneficiary identifier [BENE_ID], the claim identifier [CLM_ID], the claim line/record number [CLM_LINE_NUM]). The CCW team uses the last date on the claim, referred to as the CLM_THRU_DT, to partition the claims into calendar year files.

Researchers may wish to obtain claims data for a population they identify with the MBISG file. Remember that the claims files will consist of either FFS, PDE, or MA encounter data, and the MBISG cohort includes all Medicare beneficiaries enrolled at a point in time (March 1, 2023, or March 1, 2024). There are instances where beneficiaries with claims or MA encounter data that do not match the MBISG and vice versa. If interested in claims for a beneficiary population, they should use the BENE_ID to perform this linkage. Once the user has identified their study population, the MBISG probability weights work as intended whether the population includes FFS-enrolled, MA-enrolled, or other Medicare population subsamples.

5.0 Receiving CCW Data

This section describes the content and format of the CCW MBISG data package researchers receive. The CCW team provides data files to the researcher in the following formats.

5.1 Within the CCW Virtual Research Data Center (VRDC)

Researchers approved for a cohort of beneficiaries extracted from the MBISG files will receive this custom cohort in their IN0nnnnn folder (where the 0nnnnn represents the researcher's DUA number). The CCW team names the file:

MBISG_MARCHSNPSHT{YY}_R99999.sas7bdat (where YY is the year of the MBISG file and 99999 represents the request number).

The data file is also a 100% pre-extracted file located in the CCW VRDC SAS library, MBISG_R. The CCW team names this data file:

MBISG_MARCHSNAPSHOT_2023.sas7bdat (and the 2024 file named MBISG_MARCHSNAPSHOT_2024.sas7bdat)

5.2 Physical Shipment of Data

Some researchers receive a physical data shipment from the CCW team. There are one or more folders on the physical media, each containing multiple files. CCW organizes the folders by request number as depicted below:

- 📁 XXXXX (folder with your CCW data request number)
- 📁 Extract file documentation

The researcher will find a year folder (e.g., 2023 or 2024) inside the request number folder that contains a Read Me file and the MBISG data file in the format of a password-protected executable files (self-decrypting archives [SDA]). If the data request contains additional types of data besides MBISG, there could be additional SDAs.

Inside the year folder, there is a Read Me file and the MBISG SDA (ex. For 2023; reference ResXXXXXXXXreqXXXXXX_2023_MBISG_MARCHSNPSHT.exe [Table 3](#) and [Table 5](#)). The naming convention for the SDA is as follows:

res<XXXXXXXXXX>req<XXXXXX>_<YYYY>_<FTYPE>

Researcher DUA#
Year of data
File type

CCW request #

For example, if the DUA number was 000077777, the CCW request number was 012345, the year will be the year of the data requested (ex. 2023) and the file type is MBISG_MARCHSNPSHT.

The folders and data files would look like this:

- 📁 12345
 - 📁 2023
 - READ_ME_FIRST_REQ12345_2023.txt
 - ResXXXXXXXXreqXXXXXX_2023_MBISG_MARCHSNPSHT.exe

Table 3. Format and naming convention for the CCW files


File	File description
READ_ME_FIRST_REQ12345_2023.txt	This is a text file that describes the files contained in the output package. Filename example: READ_ME_FIRST_REQ12345_2023.txt
res000077777req012345_2023_MBISG_MARCHSNPSHT.exe	This is the SDA executable that researchers must run to decrypt and uncompress the MBISG data file. In this example, 000077777 is the DUA number, 012345 is the request number, and 2023 is the year of the data. This executable includes the v8 SAS read-in program, the .dat file, and the .fts file containing the layout and record counts.

Table 4. MBISG SDA contents

File	File description
mbisg_marchsnapshot_resXXXXXXXXX_reqXXXXXX_2023.dat mbisg_marchsnapshot_res<0000nnnnn>_req<0nnnnn>_2023.fts mbisg_marchsnapshot_read_v8.sas	This set of files includes the MBISG .dat (data) file, .fts (layout and record counts) file, and version 8 SAS read-in program.

In addition to the specific data files the researcher requested, the CCW team includes a decryption resource file in the deliverable package. [Table 5](#) shows this file.

Table 5. CCW resources accompanying data files

File	Description
 Decryption instructions.pdf	This document contains instructions for decrypting/uncompressing the data files.

The encryption technique for files extracted from the CCW uses Pretty Good Privacy (PGP) Command Line software. This method builds a compressed, encrypted, password-protected file using a FIPS 140-1/140-2 approved AES256 cipher algorithm. The CCW team builds the SDA on the CCW production server, downloads it to a desktop PC, and burns it to a CD, DVD, or USB hard drive, depending on the size of the files.

After the CCW team ships the data to the researcher, they email the password to decrypt the archive to the researcher via email. Each researcher's request has a unique encryption. The CCW team never packages the password and the data media together. To decrypt the data files, the researcher accesses the email containing the decryption password. The data package contains detailed instructions for using this password.

6.0 Where to Get Assistance

Researchers interested in working with CCW data should contact ResDAC. They offer free assistance to researchers using Medicare data for research. The ResDAC website provides links to descriptions of the CMS data available, request procedures, supporting documentation, such as record layouts and SAS input statements, workshops on how to use Medicare data, and other helpful resources. Visit the ResDAC website at <http://www.resdac.org> for additional information.

ResDAC is a CMS contractor, and researchers should first submit requests to ResDAC for assistance in the application, obtaining, or using the CCW data. Researchers can reach ResDAC by phone at 1-888-973-7322, email at DataRequests@cms.hhs.gov, or online at <http://www.resdac.org>.

If a ResDAC technical advisor is unable to answer questions, the advisor directs the researcher to the appropriate person. If the researcher requires additional CMS data (data not available from the CCW) to meet research objectives, or has any questions about other data sources, the researcher should first visit the ResDAC website.

The CCW Help Desk staff provides assistance between 8:00 am to 5:00 pm ET, Monday through Friday (excluding most federal holidays). Contact the CCW Help Desk at ccwhelp@ccwdata.org or 1-866-766-1915.

Appendix A — List of Acronyms

Acronym	Definition
AA	Asian American
AI	American Indian
AN	Alaska Native
CCW	Chronic Conditions Warehouse
CME	Common Medicare Environment
CMS	Centers for Medicare & Medicaid Services
DUA	Data Use Agreement
ESRD	End-stage renal disease
FFS	Fee-for-service
HEDIS	Healthcare Effectiveness Data and Information Set
MA	Medicare Advantage
MAO	Medicare Advantage Organizations
MBI	Medicare beneficiary identifier
MBISG	Medicare Bayesian Improved Surname Geocoding
MBSF	Master Beneficiary Summary File
NDI	National Death Index
NHPI	Native Hawaiian or Other Pacific Islander
OMH	CMS Office of Minority Health
OTCC	Other Chronic or Potentially Disabling Conditions
PGP	Pretty Good Privacy
PII	Personally identifiable information
R/E	Race and ethnicity
ResDAC	Research Data Assistance Center
SDA	Self-decrypting archive
SSA	Social Security Administration
VRDC	Virtual Research Data Center