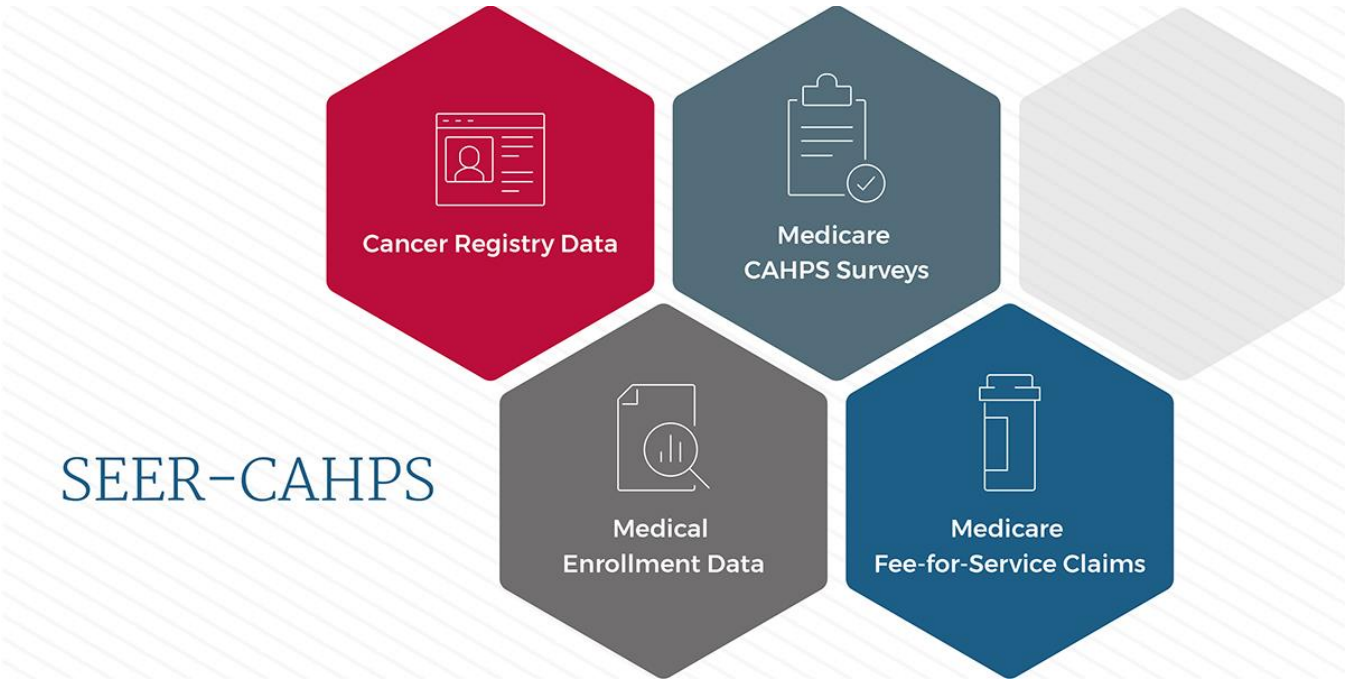
SEER-CAHPS User Guide



Draft
Version 1.3
June 2023

Contents, Tables, and Figures

1. Introduction

This user guide provides researchers with a compilation of guidance and information on the SEER-CAHPS data resource, a linkage between the National Cancer Institute's (NCI) Surveillance, Epidemiology and End Results (SEER) cancer registry data and the Centers for Medicare & Medicaid Services' (CMS) Medicare Consumer Assessment of Healthcare Providers and Systems (CAHPS®) patient surveys.

This resource, a result of collaborative effort among NCI, the SEER registries, and CMS, first became publicly available in 2016. The linkages between the different components of the SEER-CAHPS data are updated every 2-3 years, based on SEER and other data availability.

It's important to note that the most recent data linkage was completed in 2022. It includes SEER data for individuals diagnosed with cancer in 1975-2019, CAHPS data for 1997-2019, and Medicare claims data for Fee-for-Service beneficiaries for 1999-2019. This guide reflects the 2022 linkage and data elements, although some portions pertain to earlier linkages. Updates will occur with future linkages. The most up-to-date guide can be found on the SEER-CAHPS website here: https://healthcaredelivery.cancer.gov/seer-cahps/researchers/guidance.html.

1.1 Need Additional Help?

For more information or additional guidance on SEER-CAHPS data, please contact the SEER-CAHPS staff using this online form: https://healthcaredelivery.cancer.gov/seer-cahps/contact.html or at the following email address: NCISEERCAHPS@mail.nih.gov.

2. SEER-CAHPS Basics

SEER-CAHPS is a linked data resource for research on the quality of cancer care. These data provide a rich opportunity for analyses of Medicare beneficiaries' experiences with their care at various points on the cancer care continuum.

Research using SEER-CAHPS data have the potential to fill an important gap in existing knowledge by enabling comparisons of patients' care experiences between MA and FFS beneficiaries and between patients with and without cancer. For Medicare FFS beneficiaries, the SEER-CAHPS data set also allows for the evaluation of their health care utilization and costs of care through the linkage to Medicare claims.

Table 1 below provides an overview of what types of data come from each part of the linkage.

Table 1. What types of data come from each part of the linkage?

Variables	SEER	Medicare Claims and Enrollment	CAHPS
Cancer Site/Stage	x		
First course of treatment (radiation/surgery)	x		
Cause of death	x		
Vital status	x	x	
Cost of care & service utilization		x	
Claims (for Fee-for-Service enrollees)		x	

Variables	SEER	Medicare Claims and Enrollment	CAHPS
Global ratings of care			x
Care composites			x
Health status			x
Patient demographics	x	x	x

2.1 About SEER

The SEER Program works to provide information on cancer statistics in an effort to reduce the burden of cancer among the U.S. population. SEER started collecting data on cancer cases in 1973 with a limited number of registries and continues to expand to include even more areas. You can learn more about the SEER data in **Sections 4.4, Identifying Cancer Information: Diagnoses, Site, and Stage**, and **5.1.1, SEER Cancer Registry Data**.

2.2 About CAHPS

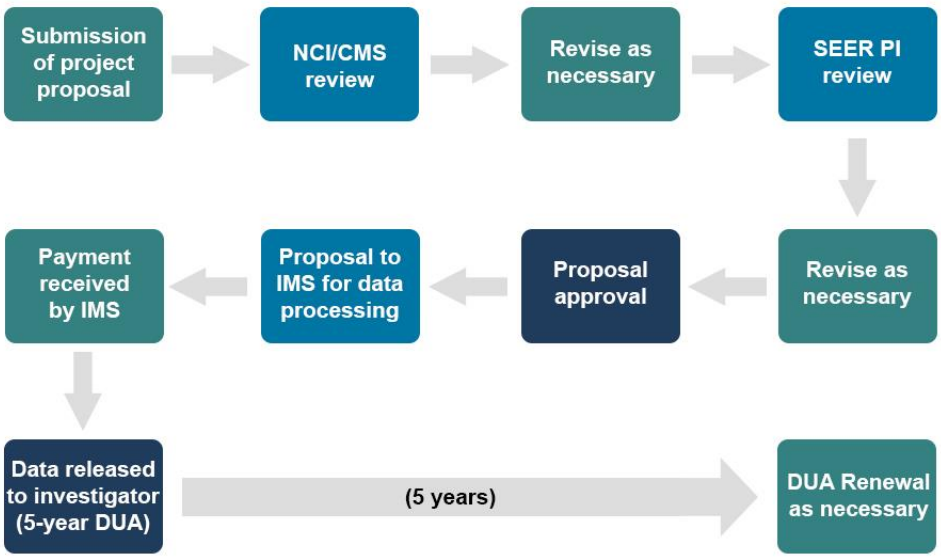
Since 1997, CMS has sponsored annual administrations of the Medicare CAHPS surveys to assess the health care experiences of Medicare enrollees in Medicare Advantage (MA) and fee-for-service (FFS) plans. The CAHPS surveys are widely used instruments for measuring US health care quality from patients' perspectives.

It is important to note that the CAHPS survey is not designed to be longitudinal. However, a small proportion of respondents in SEER-CAHPS may have completed more than one CAHPS survey.

3. Obtaining the Data

The SEER-CAHPS linked data are available to outside investigators for research purposes. Although personal identifiers for all patient and medical care providers have been removed from the SEER-CAHPS data, there remains the remote risk of re-identification (given the large amount of data available). In light of the sensitive nature of the data, maintaining patient and provider confidentiality is a primary concern of the NCI, SEER, and CMS. Therefore, the SEER-CAHPS data are not public use data files. Investigators are required to

Figure 1. Obtaining SEER-CAHPS Data



obtain NCI approval in order to obtain the data. Approval or exemption by an institutional review board is also required.

An application form and data use agreement (DUA) form can be found on the SEER-CAHPS website: https://healthcaredelivery.cancer.gov/seer-cahps/obtain/required.html. Submissions require a cover letter, application and DUA form, and Institutional Review Board (IRB) approval. We strongly recommend that investigators schedule a phone call with the SEER-CAHPS team prior to submitting their draft proposal.

Representatives from NCI, CMS, and SEER will review each proposal. The review and approval process generally takes 4-6 weeks from initial receipt of proposals. This is an iterative process with multiple steps, as shown in **Figure 1**.

3.1 Process for Obtaining Data after Approval

Once a data request has been approved and all required documents are on file, IMS (NCI's programming contractor) will provide an invoice to the investigator to cover the costs of creating the requested data files (see Cost of Acquiring SEER-CAHPS Data). The SEER-CAHPS website has a calculator to see how much the data you seek will cost: https://healthcaredelivery.cancer.gov/seer-cahps/obtain/costcalc.html

In accordance with an NCI-IMS contractual agreement, IMS will begin processing data requests upon receipt of payment. IMS requires pre-payment of all invoices.

See **Section 5, About the Data** , for further information on how the data are delivered and set up for analysis.

3.2 Data Policies

3.2.1 Data Updates

The 2022 SEER-CAHPS linkage incorporates claims data from the Chronic Conditions Warehouse (CCW) and is not compatible or linkable with any previous release. Therefore, you will not be able to request updated data for any application approved prior to this release. If you want data from the 2022 SEER-CAHPS linkage, you will need to submit a new application. All previously released data must be destroyed.

For applications that are approved for the 2022 linkage, you will be able to request data up to three times: the initial data request and then updated data from the next two subsequent linkages. If additional updates are desired, investigators will need to submit a new application for review and approval.

3.2.2 Data Retention

The Data Use Agreement (DUA) states the data retention time period is 5 years. If additional time is necessary to complete the approved project, investigators must request a one-year extension to the DUA. These extensions may be renewed annually until a maximum data retention period of 10 years. If more

than 10 years has lapsed since data were initially received, investigators will need to submit a new application for review and approval. Without an approved extension, all SEER-CAHPS data must be destroyed.

3.2.3 Data Sharing

Investigators will be allowed to share data for approved projects (see Obtaining the Data) with colleagues at their institute only if the data:

- pertain to the same cohort (e.g., the same cancer site), and
- were purchased within the previous 2 years.

Please note that the data retention period for the shared data will commence from when the data for the initial project were received, not when the request to share the data was submitted.

3.2.4 Data Usage Agreement Amendments

Investigators wishing to make changes, including the addition of a study aim, to an active DUA without requesting additional data must provide written documentation pertaining to such modifications. NCI will review and approve proposed amendments on a case-by-case basis. All potential DUA amendments are subject to the same provisions specified in the SEER-CAHPS application, including:

- Additional aims and/or content closely relate to aims found in the original proposal.
- Proposal additions are relevant to improving the quality of care of older cancer patients.

Amended proposals can be submitted for review via e-mail to NCISEERCAHPS@mail.nih.gov.

3.3 Frequently Asked Questions

Is there a limit to the number of cancer sites that a researcher may request?

In order to balance the preferences of our investigators with our charge to be good stewards of the linked data resource, we must limit the number of cancer sites for all data use agreements (DUAs). Written justification (based on study rationale and existing literature) is needed for all cancer sites requested. Furthermore, we are willing to revisit the release of additional cancer sites upon demonstration of publication on a smaller subset of cancer sites.

What criteria are used to review data use agreements and proposals?

We assess proposals in order to ensure data safety and confidentiality, but we also keep in mind the ability of the dataset to meet proposed aims. Investigators are advised to carefully consider each cancer site they are requesting and give a rationale for each one. We also suggest that investigators review previously published manuscripts using SEER-CAHPS data so as not to duplicate previously published work.

Do only the Principal Investigators need to submit a signed DUA?

All personnel with access to the SEER-CAHPS data should provide signatures on the DUA.

In general, how long will it take to receive data once the proposal is completed and approved by the NCI and SEER PIs?

The entire process, including submission, multilevel review, feasibility checks, invoicing, and data delivery, usually takes 3-4 months.

As part of the approval process, does NCI critique the methodology or merits of the proposed projects?

The purpose of the approval process is to ensure the confidentiality of the patients and providers in the SEER geographic areas. If there are concerns about confidentiality, SEER-CAHPS data will not be released, regardless of whether a researcher has already been funded by another agency or organization to conduct an analysis using the data. Reviewers from NCI and SEER *may* comment, however, on aspects of the research plan that may affect project feasibility and scientific rigor. NCI will work with investigators requesting data files to balance their research needs with those of the individuals and institutions included in the data.

Can data on restricted variables be requested in a project proposal?

If investigators determine that restricted variables (such as unencrypted physician identifiers) are an essential part of the analysis, data for these variables will be available upon request and evaluated on a case-by-case basis. Investigators intending to include restricted variables in their proposals must include detailed justification for access to the restricted variable(s). For additional information, please contact SEER-CAHPS staff via email at NCISEERCAHPS@mail.nih.gov.

If I already have applied for and obtained SEER-Medicare data, do I only need to pay the cost of adding the CAHPS data?

The SEER-CAHPS data are a different linkage than SEER-Medicare, and are based upon a different sampling frame, which is those who complete a CAHPS survey. As a result, a researcher cannot add the CAHPS survey data to previously obtained SEER-Medicare data. The cost of SEER-CAHPS is also separate from the cost that you may have paid for SEER-Medicare data. SEER-CAHPS data files are created by Information Management Services (IMS), and the cost of data reimburses IMS for the cost of producing the data.

One of the required documents is evidence of IRB approval or exemption. Can I submit the DUA application for review and then follow up with the IRB approval/exemption before the data are released to me?

IRB approval documentation is required for application to obtain the SEER-CAHPS data. A researcher can submit the DUA application for review and then follow-up with the IRB approval or exemption, however no data will be released before we receive the IRB approval/exemption.

How is the grant submission process different from the application process to receive the data?

The grant proposal and DUA proposal are separate processes, as the grant proposal is to apply for funding, and the DUA proposal is the process to obtain the data. We encourage investigators interested in the SEER-CAHPS data resource, however, to reach out to SEER-CAHPS staff prior to submitting a grant proposal for a project requiring SEER-CAHPS data.

Is SEER-CAHPS the SEER-Medicare linkage with the CAHPS survey added?

The SEER-CAHPS data are a different linkage than SEER-Medicare, and are based upon a different sampling frame, which is those who complete a CAHPS survey. Please see **Table 2** for more on how SEER-CAHPS differs from the SEER-Medicare and the SEER-Medicare Health Outcomes Survey (MHOS) linkages.

Table 2. Differences between SEER-CAHPS and other linked SEER data resources

	SEER CAHPS	SEER MHOS	SEER Medicare
SEER Cancer Registry Data	✓	✓	✓
Medicare Enrollment Data	✓	✓	✓
Medicare Advantage Enrollees	✓	✓	✓
Fee-for-Service Enrollees	✓	✗	✓
Claims Data	✓	✗	✓
Part D Claims Data	✓	✓	✓
Physician & Hospital Characteristics	✓	✗	✓

	SEER CAHPS	SEER MHOS	SEER Medicare
CAHPS Experience of Care Survey Data	✓	✗	✗
MHOS Quality of Life Survey Data	✗	✓	✗
MDS and OASIS Assessment Data	✓	✗	✓

3.4 Tutorials and Other Support

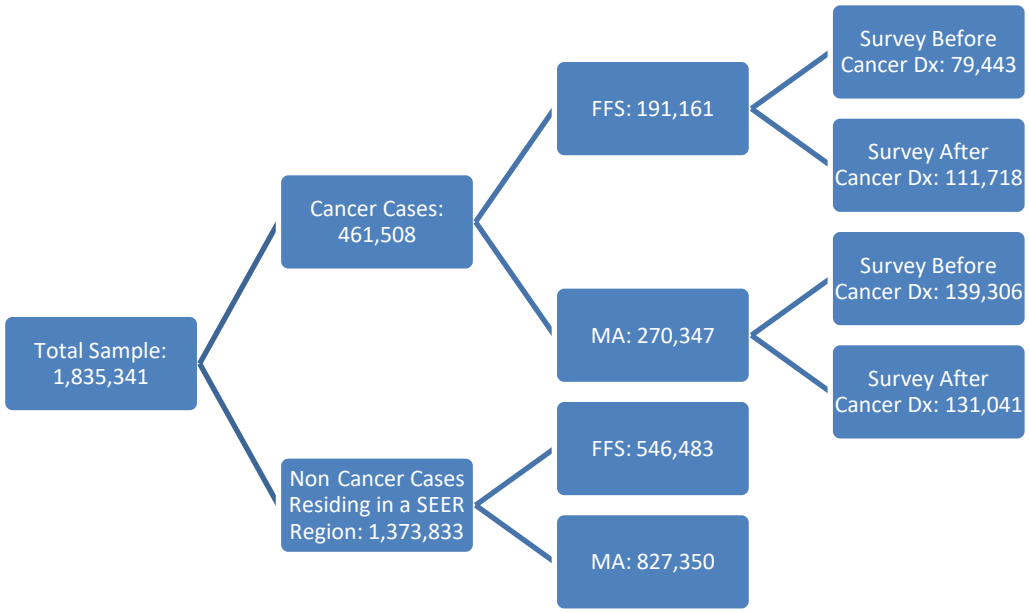
Tutorials and webinars on the SEER-CAHPS data are available on the website: https://healthcaredelivery.cancer.gov/seer-cahps/researchers/

Analytic guidance for researchers, including the use of survey weights, case-mix adjustment, analytic approaches, missing data, and response patterns are available in this User Guide and online: https://healthcaredelivery.cancer.gov/seer-cahps/researchers/guidance.html

4. Who Is in SEER-CAHPS?

SEER-CAHPS includes all Medicare Advantage and Fee-for-Service beneficiaries who completed a CAHPS survey while living in a SEER region. Below is a figure that shows the number of CAHPS survey respondents in SEER-CAHPS based on cancer diagnosis in a SEER region and Medicare plan type.

Figure 2. Sample characteristics: enrollment type and survey timing, SEER-CAHPS 1997-2019



Dx: diagnosis; FFS: fee-for-service; MA: Medicare Advantage; SEER: Surveillance, Epidemiology, and End Results

4.1 People With and Without Cancer in SEER Areas

Refer to **Section 4.4, Identifying Cancer Information: Diagnoses, Site, and Stage**, for information on people with cancer.

The non-cancer sample available to researchers only includes people residing in SEER areas at the time of their survey, since people outside SEER areas have uncertain cancer histories. Although there is self-reported information on whether someone has ever been diagnosed with cancer other than skin cancer (CND_CANCER), the responses are sometimes inconsistent or unreliable. That is, it is possible that individuals in the "non-cancer" population may have been diagnosed with cancer prior to residing in a SEER region.

In the CAHPS survey metadata, the variable CA_STAT is coded as follows:

- 1=Non-melanoma (skin) cancer before survey in SEER. If same year, but month unknown, "any cancer" question (CND_CANCER) must = yes or missing.
- 2=No cancer including skin before survey in SEER and "any cancer" question = no or missing. If same year but month unknown, "any cancer" must = no.
- 3=No cancer before survey in SEER and "any cancer" question= yes.
- 4=Not in SEER, resided in SEER area and "any cancer" question = yes
- 5=Not in SEER, resided in SEER area and "any cancer" question = no or missing
- 6=Not in SEER and did not reside in SEER area
- 7=Melanoma cancer in SEER before survey or in same year as survey, month is same or missing
- 8=Non-malignant tumors before the survey, or in same year as survey, month same or missing
- 99=Not classified

This variable can be used to define a cohort that, based on multiple data sources, has had no known cancer diagnoses (i.e., CA_STAT = 2).

4.2 Patient_ID: The Unique Identifier

The variable PATIENT_ID is an encrypted identifier that protects privacy while still allowing us to link unique individuals across different years and types of data. Note: earlier linkages cannot be linked to the current linkage using a unique ID. This number does not change during a beneficiary's lifetime, and the Chronic Conditions Warehouse (CCW) uses each number only once. The PATIENT_ID is specific to the CCW and is not applicable to any other identification system or data source.

4.3 Demographic Characteristics: Age, Sex, Race, Living Arrangements

Demographic characteristics are available from multiple sources. For example, there are:

- Age as of survey response from the CAHPS data:
    - Computed age on Nov 30 of survey year from SA_DOBSAS (SA_AGECALC)

    - A 9-level categorical variable (based on computed age) that collapses ages 65-74 (AGE9)
    - A 10-level categorical variable that reports ages 65-74 in 5-year groups (AGE10)
    - Age category (Used computed age when age question missing) – a 7-level categorical variable that collapses ages 18-44 (AGECAT)
- Age as of diagnosis date from the SEER data (AGER1-AGER10)
- Actual dates of birth from the Medicare data (BENE_BIRTH_DT)

Deciding on which of the age variables to use will depend on the goals of the analysis.

Similarly, analysts will find multiple sources of data on race and ethnicity:

- Race variables derived from the CAHPS survey data
    - A constructed variable with 7 *mutually exclusive* categories derived from the CAHPS responses on Hispanic ethnicity and race categories combined (SC_RACE)
- Six individual *non-exclusive* race/ethnicity binary (0/1) flags from the CAHPS responses:
    - White (RACE_WHITE)
    - Black (RACE_BLACK)
    - Asian (RACE_ASIAN)
    - Native Hawaiian/Pacific Islander (RACE_PACIFIC)
    - Native American (RACE_NATAMER)
    - Hispanic (RACE_HISP)
- Race variables from the Medicare Beneficiary Summary File (MBSF)
    - RACE - a 7-level categorical variable derived from Medicare and Social Security data
    - RTI_RACE_CD – a 7-level categorical variable derived from Medicare and Social Security data that employs an algorithm based on first and last names that may be Hispanic or Asian in origin
- Race variables from the SEER data
    - Consult the SEER documentation

The choice of which race variable to use is up to the investigator; however, we advise researchers to consider using self-reported race/ethnicity (SC_RACE) supplemented with SEER or Medicare's race variable where needed.

Table 3 below provides selected sociodemographic information for SEER-CAHPS respondents in the most recent linkage.

Table 3. Selected sociodemographic and health characteristics, SEER-CAHPS 2020 Linkage

Characteristic	Cancer (n=461,508)					Non Cancer* (n=1,373,833)				
	Total	Medicare Advantage		Fee for Service		Total	Medicare Advantage		Fee for Service	
		N	%	N	%		N	%	N	%
Total	461,508	270,347	59	191,162	41	1,373,833	827,350	60	546,483	40
Age at Survey										
Under 65	30,304	16,254	6	14,050	7	168,942	94,602	11	74,340	14
65-74	227,718	138,974	51	88,744	46	671,032	415,874	50	255,158	47
75-84	159,046	91,847	34	67,199	35	401,007	243,326	29	157,681	29
85+	44,440	23,272	9	21,168	11	132,852	73,548	9	59,304	11
Gender										
Male	228,113	132,863	49	95,250	50	557,112	330,310	40	226,802	42
Female	233,395	137,484	51	95,911	50	816,721	497,040	60	319,681	59
Race/Ethnicity										
NH White	346,338	194,193	72	152,145	80	939,990	532,027	64	407,963	75
NH Black	31,923	21,402	8	10,521	6	110,112	75,221	9	34,891	6
NH Asian	17,340	11,617	4	5,723	3	71,574	48,404	6	23,170	4
NH North American Native	1,440	807	0	633	0	5,840	3,081	0	2,759	1
NH Mixed	7,396	4,293	2	3,103	2	25,682	15,337	2	10,345	2
NH Other	2,438	1,801	1	637	0	7,692	5,508	1	2,184	0
Hispanic, any race	28,671	20,513	8	8,158	4	131,807	96,383	12	35,424	6
Unknown	25,962	15,721	6	10,241	5	81,136	51,389	6	29,747	5
Education										
Less than High School	88,827	58,470	22	30,357	16	291,583	195,853	24	95,730	18
High School Graduate or GED	136,922	82,560	31	54,362	28	405,343	248,513	30	156,830	29
Some College/2-years Degree	104,678	60,860	23	43,818	23	310,397	183,794	22	126,603	23
4-years College Graduate	44,485	23,436	9	21,049	11	121,180	65,949	8	55,231	10
More than 4-years College Degree	59,267	28,056	10	31,211	16	156,926	75,883	9	81,043	15
Unknown	27,329	16,965	6	10,364	5	88,404	57,358	7	31,046	6
Smoking History										
Non-Smoker or Former Smoker	378,442	224,030	83	154,412	81	1,125,263	690,736	83	434,527	80
Current Smoker	49,711	30,680	11	19,031	10	138,208	84,939	10	53,269	10
Unknown	33,355	15,637	6	17,718	9	110,362	51,675	6	58,687	11
Survey Language										
English	435,756	254,451	94	181,305	95	1,287,318	768,851	93	518,467	95
Spanish	5,671	4,398	2	1,273	1	33,780	27,675	3	6,105	1
None/Unknown	20,081	11,498	4	8,583	4	52,735	30,824	4	21,911	4
Proxy Status										
Proxy	46,226	26,232	10	19,994	10	170,700	101,227	12	69,473	13
No Proxy	355,887	214,601	79	141,286	74	1,019,318	633,732	77	385,586	71
Unknown	59,395	29,514	11	29,881	16	183,815	92,391	11	91,424	17

Characteristic	Cancer (n=461,508)					Non Cancer* (n=1,373,833)				
	Total	Medicare Advantage		Fee for Service		Total	Medicare Advantage		Fee for Service	
		N	%	N	%		N	%	N	%
General Health status										
Excellent	33,267	20,370	8	12,897	7	117,915	72,244	9	45,671	8
Very Good	114,653	67,961	25	46,692	24	349,329	207,930	25	141,399	26
Good	168,824	100,018	37	68,806	36	471,537	288,336	35	183,201	34
Fair	101,733	58,253	22	43,480	23	298,287	179,689	22	118,598	22
Poor	26,672	13,504	5	13,168	7	82,405	43,325	5	39,080	7
Unknown	16,359	10,241	4	6,118	3	54,360	35,826	4	18,534	3

*Includes Medicare CAHPS respondents living in SEER areas who do not have any recorded cancer history

4.4 Identifying Cancer Information: Diagnoses, Site, and Stage

The SEER cancer registry data includes information about all primary cancers that a person may develop, including patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and first course of treatment, and they follow up with patients for vital status. Please see the NCI SEER-CAHPS site for more details on documentation and variables. Applications for data use agreements should specify the cancer sites included in the project, with a limit of 10 cancer sites per proposal. Some sites are generally combined and analyzed as a group, including:

- Head and neck
- Colon and rectal
- Lung and bronchus

Additional information available includes:

- Primary site (`SITE1-SITE10`)
- Stage and extent of disease (EOD)
  - AJCC 6th T, N, M and stage are available for cases diagnosed in 2004+.
  - AJCC 7th T, N, M, and stage are available for 2010+ cases.
  - If all you need is local, regional, or distant, the SEER COMBINED SUMMARY STAGE 2000 variable has fewer observations with missing information than the SEER SUMMARY STAGE 2000 version
  - Data before 2004 had a more simplified version of extent of disease than CS. An on-going SEER project is to apply AJCC 6th criteria to the earlier data in order to create longer term trends such as AJCC 6th stage for cases diagnosed in 1988 and later.

- An overview of the stage data and data submissions in SEER*Stat can be found at http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage/ .
- Major groups of histologies/behaviors were not collected consistently over time; for example, benign brain, myelodysplastic syndromes, and borderline tumors of the ovary. Researchers are advised to look carefully at the BEHTREND variables if studying any of these types of cancers. Additionally, consult histology recodes for brain cancer groupings and Ann Arbor staging for lymphomas.
- SEER, NAACR (HISTREC), ICD-O-2, and ICD-10 codes are all provided

- Laterality (LAT1-LAT10), sequence (SEQ1-SEQ10) and record # (REC01-RECNN): these variables provide diagnostic information for up to 10 diagnoses per person
- Site-specific factors:
  - For example, ER/PR status for breast cancer patients, genetic information, Gleason score for prostate cancer patients, WHO/ISUP grade
  - Site-specific Factors information from SEER

SEER Coding & Staging Manuals - codes and coding instructions for SEER data and extent of disease.

Refer to the 5-digit Site Recode Dictionary for how to identify specific cancer sites. All SEER variables copied directly from the SEER file are described in the SEER Research Data Record Description.

Tables 4 and 5 provides information about number of CAHPS survey respondents in SEER-CAHPS by selected first cancer site and time between diagnosis and survey (reported separately by enrollment type). Individuals in SEER-CAHPS who do not have a cancer diagnosis are not included in this table. The cancer sites are listed in order of data frequency.

Table 4. Number of respondents by selected* first cancer site and date of diagnosis: Medicare Advantage

First cancer	Total number of SEER Linked patients	First survey before month of first cancer diagnosis		First survey within 2 years of first cancer diagnosis		First survey within 3 5 years of first cancer diagnosis		First survey within 6 10 years of first cancer diagnosis		First survey within 11+ years of first cancer diagnosis	
		N	%	N	%	N	%	N	%	N	%
Prostate	48,758	16,871	35	7,423	15	7,411	15	10,062	21	6,991	14
Breast	47,047	17,646	38	6,144	13	5,943	13	8,098	17	9,216	20
Colorectal	28,023	14,343	51	3,529	13	2,904	10	3,628	13	3,619	13
Lung and Bronchus	27,438	21,719	79	2,793	10	1,268	5	1,023	4	635	2

First cancer	Total number of SEER Linked patients	First survey before month of first cancer diagnosis		First survey within 2 years of first cancer diagnosis		First survey within 3 5 years of first cancer diagnosis		First survey within 6 10 years of first cancer diagnosis		First survey within 11+ years of first cancer diagnosis	
		N	%	N	%	N	%	N	%	N	%
Melanoma Skin	15,399	7,269	47	1,946	13	1,867	12	2,136	14	2,181	14
Bladder	14,782	8,241	56	1,889	13	1,537	10	1,674	11	1,441	10
Non-Hodgkin Lymphoma	9,905	5,649	57	1,231	12	1,014	10	1,114	11	897	9
Uterine Corpus	8,250	2,986	36	961	12	999	12	1,339	16	1,965	24
Kidney/Renal Pelvis	7,084	3,740	53	939	13	838	12	901	13	666	9
Head/Neck	6,718	3,298	49	923	14	727	11	902	13	868	13
Leukemia	5,992	3,776	63	838	14	486	8	561	9	331	6
Pancreas	5,437	4,890	90	352	6	99	2	63	1	33	1
Stomach	3,448	2,424	70	411	12	202	6	225	7	186	5
Ovary	2,913	1,633	56	354	12	233	8	252	9	441	15
Liver/Bile Duct	2,650	2,138	81	265	10	120	5	88	3	39	1
Esophagus	1,863	1,438	77	193	10	105	6	88	5	39	2
Uterine Cervix	1,056	335	32	87	8	97	9	157	15	380	36

* Sites reflect most common cancer sites in SEER-CAHPS.

Table 5. Number of respondents by selected* first cancer site and date of diagnosis: Fee-For-Service

First cancer	Total number of SEER Linked patients	First survey before month of first cancer diagnosis		First survey within 2 years of first cancer diagnosis		First survey within 3 5 years of first cancer diagnosis		First survey within 6 10 years of first cancer diagnosis		First survey within 11+ years of first cancer diagnosis	
		N	%	N	%	N	%	N	%	N	%
Prostate	35,812	9,395	26	5,206	15	5,785	16	8,582	24	6,844	19
Breast	34,571	10,286	30	4,361	13	4,982	14	7,026	20	7,916	23
Colorectal	17,915	7,109	40	2,565	14	2,379	13	3,011	17	2,851	16
Lung and Bronchus	17,245	12,241	71	2,187	13	1,210	7	1,064	6	543	3
Melanoma Skin	13,634	5,016	37	1,918	14	1,956	14	2,375	17	2,369	17
Bladder	10,348	4,511	44	1,549	15	1,394	13	1,617	16	1,277	12
Non-Hodgkin Lymphoma	7,244	3,247	45	1,038	14	979	14	1,149	16	831	11
Uterine Corpus	6,016	1,798	30	731	12	787	13	1,151	19	1,549	26
Kidney/Renal Pelvis	5,284	2,302	44	746	14	726	14	869	16	641	12
Head/Neck	4,834	1,970	41	694	14	671	14	788	16	711	15
Leukemia	4,532	2,345	52	656	14	523	12	623	14	385	9
Pancreas	3,267	2,792	85	280	9	84	3	71	2	40	1
Ovary	2,046	938	46	290	14	223	11	241	12	354	17
Stomach	2,006	1,271	63	243	12	161	8	192	10	139	7
Liver/Bile Duct	1,534	1,139	74	193	13	98	6	79	5	25	2
Esophagus	1,173	798	68	161	14	90	8	77	7	47	4
Uterine Cervix	688	170	25	71	10	71	10	122	18	254	37

* Sites reflect most common cancer sites in SEER-CAHPS.

4.4.1 Malignant and Benign Tumors

Individuals diagnosed between 1999 and 2019 with both malignant and benign tumors are in the data. The SEER variables FRSTPRM1-FIRSTPRM10 can be used to exclude benign tumor diagnoses. Refer to the SEER-Medicare documentation for full details.

The following example SAS code creates an inclusion indicator (INCL1) that is 1 if a person had only one primary cancer diagnosis and that cancer was malignant and 0 otherwise:

```
IF NUMPRIMS=1 AND FRSTPRM1=1 THEN INCL1=1;
ELSE INCL1=0;
```

4.4.2 Non-Melanoma and Melanoma Skin Cancers

Non-melanoma skin cancers (NMSCs; e.g., basal and squamous cell carcinomas) are the most common cancers diagnosed in the US. They rarely metastasize or need intensive management beyond outpatient surgery, which is usually curative. They are not required to be reported to SEER cancer registries. However, the very helpful CA_STAT variable (described in **Section 0,**

Dx: diagnosis; FFS: fee-for-service; MA: Medicare Advantage; SEER: Surveillance, Epidemiology, and End Results

People With and Without Cancer in SEER Areas) can be used to identify individuals who have had NMSCs, and the ICD-O-3 codes provided in the SITERWHO1-SITERWHO10 also identify some cases. Researchers seeking to answer research questions regarding NMSCs using SEER-CAHPS data should acknowledge these limitations when reporting their results.

4.5 Timing of Survey Relative to Diagnosis

One of the most important questions that investigators should answer as they design their study is how the survey response date is related to the diagnosis date in terms of defining a study period. This is because the survey asks respondents to think about the past 6 months when responding.

If you are interested in understanding the associations between care experiences and specific outcomes, the longer the time elapsed between the survey date and the date of the outcome, the less certain you will be that the responses were pertinent to the outcome.

Several SEER-CAHPS variables are helpful in defining the time period, including:

- Number of cancers before survey (NUMCABEF)
- Number of cancers after survey (NUMCAAFT)
- Number of months from first cancer to survey (TMFCA2SV)
- Sequence # of most recent cancer before survey (SEQCABEF)
- Survey date received (SVY_DT_RCV)

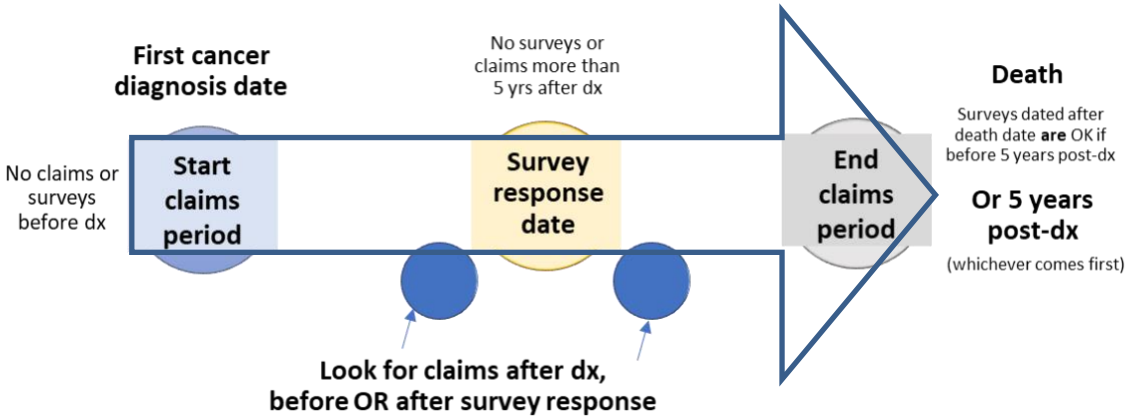
- This date is imputed in certain years – for example, although MA surveys were fielded in 2006 from January 6 through April 30, those survey responses all have svy_dt_rcv = 3/4/2006. Additionally, the FFS surveys were not fielded in 2006. See the Details for Researchers file for complete details.

Additional variables from SEER and the Medicare enrollment data will also be needed, including:

- Cancer diagnosis dates:
  - Month of diagnosis for each of up to 10 different diagnoses (`MODX1 - MODX10`)
    - Note that day of diagnosis is not provided. Many researchers assign the first day of the month as the date for purposes of calculating time since diagnosis or survival time
  - Year of diagnosis for each of up to 10 different diagnoses (`YRDX1 - YRDX10`)
- Date of death has two sources and a variable describing the degree to which they agree:
  - Year and month of death according to SEER (`SER_DODY, SER_DODM`)
  - Month, day, and year of death according to Medicare (`MED_DODM, MED_DODD, MED_DODY`)
  - A flag indicating the level of agreement between SEER and Medicare on the patient's month of death (`DOD_FLG`)

We suggest that investigators develop a graphic or illustration of the key dates to aid reviewers and others in understanding these nuances. A sample graphic is provided below.

Figure 3. Sample timeline illustration



4.6 Medicare CAHPS Survey Respondents

CMS conducts annual analyses to determine non-response patterns for Medicare CAHPS surveys. Unit response rates follow patterns typical for health surveys, including higher response rates for non-Hispanic whites than for other racial/ethnic subgroups, higher response rates through age 79, and lower response rates for low-income beneficiaries.

Though we do not have specific non-response information for SEER-CAHPS, we recommend referring to and citing relevant analyses using Medicare CAHPS and using provided weights for SEER-CAHPS analyses to account for sample design and non-response. Please find additional information on the CMS website.

Table 6 below includes information on the types of survey administered based on Medicare coverage type. Note that the **coverage at the time of the survey** dictates which survey a respondent received; this coverage can change over time. Thus, analysts should make sure that an enrollee's type of coverage is determined using the Medicare enrollment data rather than which survey was administered. See **Section 5.1.3, Medicare Enrollment Data** for additional details.

Table 6. Medicare CAHPS surveys and years

Survey	Years	Care Addressed
Fee-for-Service (FFS) Only	2000-2004; 2007-2010	All aspects of care for those with FFS only (without Part D)
FFS + Prescription Drug Plan (FFS+PDP)	2007-2010	All aspects of care for FFS+PDP enrollees
Fee-for-Service (FFS)	2011-2019	Non-part D aspects of care for FFS enrollees
Prescription Drug Plan (PDP)	2011-2019	Part D aspects of care for FFS enrollees
Medicare Advantage (MA) Only	1997-2005; 2007-2019	All aspects of care for MA enrollees without Part D
Medicare Advantage Prescription Drug Plan (MA-PD)	2007-2019	All aspects of care for MA-PD enrollees
Medicare Advantage Preferred Provider Organization (PPO)	2009-2012	All aspects of care for MA-PPO enrollees

You can find the CAHPS survey instruments on the Medicare Advantage and Prescription Drug Plan CAHPS® Survey website.

4.7 Medicare Enrollee Types

In this section, we briefly describe aspects of how Medicare is structured that can affect how the SEER-CAHPS data are analyzed and interpreted.

4.7.1 Fee-for-Service

Traditional Medicare, also known as fee-for-service (FFS), pays providers a set amount per procedure, event, or visit/stay. FFS enrollees can generally see any doctor or hospital in the US that accepts Medicare. FFS enrollees may also elect to have Medicare supplemental insurance (Medigap) that helps enrollees pay for out-of-pocket costs.

Roughly 60% of all Medicare enrollees were in FFS as of 2020. About 43% of Medicare enrollees say they evaluate their coverage options at least once every year.[1] A small share of beneficiaries voluntarily switch plans each year. Some may elect to move from MA to FFS to take advantage of the wider availability of providers who may specialize—for example, a provider who specializes in a specific cancer surgery. Medicare beneficiaries automatically move into FFS after electing hospice (the so-called "hospice carve-out").

When evaluating care experience measures dealing with a CAHPS respondent's health plan, it is important to keep in mind which type of coverage the individual had, and whether they still had that coverage during the period in which you are examining your outcome measure(s) or other predictors. If you wish to analyze FFS claims, check for continuous coverage in Parts A (inpatient) and B (outpatient) ***as well as*** no HMO indicators during any part of your claims period. Otherwise, you may miss some care provided when an enrollee was not in FFS, and thus does not have claims data available. See **Section 5.1.3, Medicare Enrollment Data** for additional details on checking Medicare eligibility by month and other information from the Medicare enrollment data.

Even having continuous FFS coverage is no guarantee that all of a beneficiary's utilization is observable. For example, if the beneficiary has not enrolled in Part D coverage, we cannot observe their prescription drug utilization any more than if that beneficiary buys prescription drugs from an online pharmacy or goes abroad for a surgery or other treatment.

Are Medicare Part D data part of the SEER-Medicare linked data resource?

As of the 2022 linkage, Medicare Part D claims data are available for all Part D enrollees (earlier linkages had Part D data only for cancer cases).

4.7.2 Fee-for-Service + Prescription Drug Plan Enrollees

Medicare Part D, which began in 2006, is a prescription drug benefit that subcontracts prescription drug coverage to private companies that offer prescription drug plans (PDPs). These vary in their coverage and benefit structures. The FFS + PDP surveys were fielded beginning in 2007 and ask respondents about their experiences with their PDP, as well as other aspects of care. Among the 46 million Part D enrollees in 2020, 20.2 million (44%) were in PDPs.[2]

4.7.3 Medicare Advantage

Medicare Advantage (MA), also known as Medicare Part C, is a managed-care option for Medicare beneficiaries. Under this option, MA enrollees sign up with a private health plan provider that subcontracts with Medicare to provide care for enrollees within a specific budget. MA enrollees get access to enhanced benefits, such as vision and dental coverage, as well as lower out-of-pocket costs, in exchange for restricted provider networks.

About 40% of all Medicare enrollees were enrolled in MA as of 2020, a substantial increase since this type of coverage was introduced in the late 1990s. About half of MA enrollees are in group plans offered by employers and unions. These proportions vary substantially by geography.[3]

Important note: MA plans must survey a representative sample of their members each year. This means that MA enrollees are oversampled in the SEER-CAHPS data. To account for their over-representation, we advise researchers to use survey analysis methods, including weights and strata, to produce nationally representative estimates. See **Section 6.6, Survey Analysis: Weights, Strata, and Methods** for additional guidance.

Since Medicare pays the insurer a fixed amount per enrollee to provide benefits covered by Medicare, claims are not available for MA beneficiaries in SEER-CAHPS. The exception is claims after enrollment in hospice care, at which point beneficiaries are automatically switched to FFS coverage.

The MA surveys were fielded beginning in 1997 and ask respondents about their experiences with their health plan, as well as other aspects of care. The MA surveys differ from the FFS surveys in several ways. Most notably, the overall rating of Medicare/plan: the FFS survey asks about Medicare, while the MA survey names the enrollee's specific MA health plan provider.

4.7.4 MA-Prescription Drug

MA-PD plans cover all Medicare benefits, including drugs. In 2020, 19.3 million (41%) beneficiaries were in MA-PD plans.[2]

4.7.5 MA-Preferred Provider Organizations

Preferred Provider Organizations (PPOs) are a type of MA plan that involves restricted networks but allow beneficiaries to see non-network providers for an additional cost. While MA-PPOs are still offered, a separate survey for MA-PPOs was only fielded in 2009-2012.

4.7.6 Dual (Medicare-Medicaid) Enrollees

Dual enrollees are Medicare enrollees with lower incomes for whom Medicaid is a secondary payer. This additional coverage reduces out-of-pocket costs for dual enrollees. Beginning in 2007, all dual enrollees were automatically enrolled in Part D.

A SEER-CAHPS paper published in 2019 found that dual enrollees were more likely than Medicare-only enrollees to report better experiences with their health plan and prescription drug plan. On other measures, they were no more or less likely to report worse care experiences.[4]

Dual enrollees can be identified in several ways:

- Medicaid dual eligible flag (1997-2005)– indicates dual enrollment at the time the individual was surveyed (`SA_MDCD_DUALFLG`)
- Constructed: Medicaid dual eligible status flag (2007-2019) – indicates dual enrollment at the time the individual was surveyed (`SC_DUAL_STATUS`)
- Constructed: Low Income Subsidy (2007-2019)– indicates that the enrollee's coverage was subsidized due to low income at the time the individual was surveyed (`SC_LIS`)
- State buy-in: a monthly flag variable within each annual MBSF file that indicates that the beneficiary received Medicaid or other state assistance to low-income individuals (`MDCR_ENTLMT_BUYIN01-MDCR_ENTLMT_BUYIN12 = C`)

No information is specifically available on full vs. partial benefits. However, the percentage of Federal poverty variable (`SA_FPL_PCT`) provides additional information on the incomes of selected respondents.

5. About the Data

Extracted files are sent in column-delimited files and SAS c-port format. In order to ensure the security of the patient's information during transition of files, the data files will be encrypted to a thumb drive that is password protected. The data files will also be compressed using the GZIP compression utility. A program will be made available to unzip the files onto the user's PC in the directory that the user specifies. The PC must be equipped with the Windows Operating system. GUNZIP is necessary to unzip the files if using a UNIX or Linux machine.

Sample input statements are available here: https://healthcaredelivery.cancer.gov/seer-cahps/support/.

5.1 Types of Data Included

SEER-CAHPS includes data from multiple sources: the SEER cancer registry data, Medicare CAHPS surveys, Medicare claims (for FFS beneficiaries), Medicare enrollment data, and assessment data from home health and skilled nursing episodes. This section provides an overview of each data source and important points to keep in mind when analyzing the data.

5.1.1 SEER Cancer Registry Data

The SEER program consists of several population-based tumor registries that capture information on all newly diagnosed cancer patients within their catchment area. SEER data include stage at diagnosis and information on cancer-directed surgery or radiation therapy as part of the first course of treatment, within 4 months of diagnosis. Registries do not include information on recurrence or metastasis subsequent to the initial diagnosis, nor is stage updated to reflect progressions.

5.1.1.1 SEER Cancer Registries Included in SEER-CAHPS

Registry participation in the SEER Program has changed over time. Therefore, the years of cancer diagnoses included in the SEER-CAHPS data varies by registry. The current SEER-CAHPS data includes persons who received a cancer diagnosis in:

- 1999-2019: California (Los Angeles, San Francisco, San Jose), Connecticut, Detroit*, Hawaii, Georgia, Iowa, New Mexico, Seattle, and Utah
- 2000-2019: California (all areas not indicated above), Idaho**, Louisiana, Kentucky, Massachusetts**, New Jersey, and New York**
- 2019: Texas

*As of 2021, Detroit is no longer in the SEER Program but is included in the current data release for all listed years.
**Idaho, Massachusetts and New York did not join the SEER Program until 2018, but limited variables for cases diagnosed during the listed years are included in the current data release.

The SEER data released as part of SEER-CAHPS are in a customized file known as the SEER-Medicare Cancer File. The Cancer File contains one record per person for individuals in the SEER database who have been matched with Medicare enrollment records. Basic SEER diagnostic information is available for up to 10 diagnosed cancer cases for each person.

Refer to the 5-digit Site Recode Dictionary for codes for each cancer site. The SEER Research Data Record Descriptions are available on the National Cancer Institute's Healthcare Delivery Research Program webpage.

5.1.2 Medicare CAHPS Survey Data

You can find the Medicare CAHPS survey instruments on the Medicare Advantage and Prescription Drug Plan CAHPS® Survey website. Refer to the Details for Researchers file for specific item availability by year, survey/population, and for changes in wording over time.

5.1.2.1 Medicare CAHPS Data Dictionary

- Data Dictionary: MS Excel format (XLSX, 100 KB)
- Data Dictionary: Portable Document Format
  - 1997 – 2005 (PDF, 561 KB)
  - 2007 – 2019 (PDF, 710 KB)

5.1.2.2 Survey Metadata: Mode, Version, Language, and Multiple Surveys

Researchers may consult several variables to understand more about the Medicare CAHPS survey(s) a respondent has completed:

- A flag for whether the survey was completed by mail or phone (`SVY_MODE`)
- Survey language (`SVY_SPAN`)
  - Responses include English, Spanish, or Chinese; Chinese language is one of the standard case-mix adjustment variables (see **Section 6.3, Covariate Adjustment**)
- Number of completed surveys since 1997 (`NUMCOMP`)
- Survey counter starting with survey in 1997 (`SRVSEQ`)

A small number of beneficiaries have completed multiple surveys. For these respondents, analysts should decide which responses to keep. Generally, the survey closest in time to the outcome measurement period (i.e., most recent) will be the most pertinent, keeping in mind that the surveys ask respondents to think about the past 6 months of care when selecting their responses.

5.1.3 Medicare Enrollment Data

The Medicare Beneficiary Summary Files (MBSF) include information on enrollees' month and year of birth, date of death (if any), sex, race, and state of residence; these files are created annually. They also contain information on Medicare eligibility, reason for Medicare entitlement, and enrollment by month for the period 1996 – 2019.

Additional information on the Medicare enrollment data is available from the SEER-Medicare program: https://healthcaredelivery.cancer.gov/seermedicare/medicare/enroll.html

The Centers for Medicare & Medicaid Services (CMS) maintains an annual Medicare Beneficiary Summary File (MBSF) that includes all Medicare beneficiaries. This file has multiple segments, as briefly described below. A more detailed explanation can be obtained here from the Centers for Medicare and Medicaid's (CMS's) Chronic Conditions Data Warehouse (CCW).

1. Base (A/B/C/D) contains information on the person's date of birth, date of death (if any), sex, race, state of residence, and monthly enrollment in Part A (inpatient), Part B (outpatient), Part C (Medicare Advantage/ managed care/ HMO) and Part D (prescription drug coverage).

2. Chronic Conditions Flags Documentation contains first occurrence date, mid-year flag, and end year flag to indicate the presence or absence of 27 conditions, based on Medicare services provided beginning in 1999. As a proxy of evidence for the presence of a condition, these flags are determined based on the presence of treatment for the conditions using claims-based algorithms that were created by CMS and are available on the Chronic Conditions Warehouse website. Because the flags are determined using claims data it is not possible to ascertain the information for beneficiaries enrolled in managed care/HMOs. This limitation also applies to newly-eligible Medicare beneficiaries who may have only a partial year of FFS coverage. Thus, in order for the flag to indicate the presence of a condition, the claims for the beneficiary must indicate treatment for that condition and the beneficiary must also have had continuous Part A/B-FFS coverage during the specified time period. It is important to note that the major objective for creating the flags was to allow for a quick, initial identification and extraction of beneficiaries with a given condition from the larger Medicare population. The flag definitions were intended to be broad, so that researchers could extract the data based on the flag definitions and then refine their specifications as needed for their specific analyses. The condition definitions were not intended to calculate population statistics.

3. Other Chronic or Potentially Disabling Conditions documentation contains first ever occurrence date and end year occurrence date for an additional 35 chronic or potentially disabling conditions not included in the above chronic conditions segment (e.g., mental health; tobacco, alcohol and drug use; developmental disorders; disability related conditions; behavioral health conditions); claims-based algorithms are available on the Chronic Conditions Warehouse website. Again, these flags are determined based on Medicare services provided beginning in 1999; therefore, the same considerations outlined above pertaining to the chronic condition flags should be taken with these flags.

4. Plan Characteristics File contains Medicare Advantage plan and Prescription Drug Plan information separated into six subfiles: base/benefit file, premium file, cost sharing tier file, service area file, special needs plans file and multi-year crosswalk file. The information in the Plan Characteristics File can be linked to the Part D Drug Event File (using contract ID and plan ID) to assess for variation in utilization and costs by plan type. Please note, plan identifiers in the data were encrypted prior to 2015; the multi-year crosswalk file allows tracking plans across time. For

more information (e.g., file layout and codebooks) please visit the Chronic Conditions Warehouse; documentation for file years 2007-2014 listed under the Medicare Part D heading and file years 2015+ under the Medicare Enrollment heading.

Given there are two cohorts of persons included in the SEER-CAHPS data (persons with and without cancer), there are two subsets of MBSF data available via SEER-CAHPS (MBSF-Cancer and MBSF-Non-SEER); the file documentation is the same. The MBSF-Non-SEER file can be used to identify controls for the persons with cancer or to develop population-based estimates of health care utilization (e.g., use of cancer tests such as PSA and mammography in the entire population). The persons in the MBSF-Non-SEER file are subjects that are not in the SEER cancer file but responded to a CAHPS survey in a SEER area.

View the MBSF Record Layouts and Codebooks (listed under Medicare Enrollment)

5.1.4 Medicare Claims Data

The Medicare claims data summarize Medicare enrollment, specific healthcare services that occurred in different settings (e.g., hospitals, physician offices, outpatient clinics), and healthcare assessments (e.g., while enrolled in nursing homes or home health care). As of 2020, the earliest claims available are from 1999. Refer to the SEER-Medicare documentation for detailed information on which years of claims data are currently available: https://healthcaredelivery.cancer.gov/seermedicare/medicare/table.html.

5.1.4.1 MedPAR: Inpatient (including Skilled Nursing and Emergency Department)

The Medicare Provider Analysis and Review (MedPAR) files include all Part A (i.e., hospital) short stay, long stay, and skilled nursing facility (SNF) bills for each calendar year. Inpatient files contain one summarized record per admission. Each record includes up to 25 diagnoses (ICD9/ICD10 diagnosis codes) and 25 procedures (ICD9/ICD10 procedure codes) provided during the hospitalization, along with dates and reimbursement amounts.

Researchers interested in only short-stay hospitalizations will need to subset the inpatient files file using the variable 'MedPAR short stay/long stay/skilled nursing facility (SNF) indicator code' located in column 106 ('S' = short stay, 'L' = long stay and 'N' = skilled nursing stay).

In almost all cases, a single inpatient file record reflects a summary of all care provided during an institutional stay. However, if the stay is long, there may be more than one claim per stay. This occurs most frequently for stays in SNFs as these often span several months. SNFs records often have no discharge date as persons remain in institutions beyond the period of Medicare coverage.

View ResDAC's MedPAR File Overview

View CCW's MedPAR Record Layout and Data Dictionary

5.1.4.2 Outpatient

The outpatient file contains Part B claims for 100 percent for each calendar year from institutional outpatient providers. Examples of institutional outpatient providers include hospital outpatient departments, rural health clinics, renal dialysis facilities, outpatient rehabilitation facilities, comprehensive outpatient rehabilitation facilities, community mental health centers. Outpatient surgeries performed in a hospital will be in the hospital outpatient file, while bills for outpatient surgeries performed in freestanding surgical centers appear in the carrier claims, not in the outpatient file.

The variable `CLAIM_ID` was created to index unique claims. The variable `REC_COUNT` is a counter that enumerates each record associated with a claim, where `REC_COUNT` = 1 is also the first revenue center in the first segment of a claim. Payment amount specific to a revenue center is available beginning in 1998.

As with the carrier data, there may be multiple records for the same date of service. Additionally, data related to each revenue center on a claim are written to a separate record. Definitions for revenue center codes may be obtained by contacting ResDAC or CMS directly.

View ResDAC's Outpatient File Overview

View CCW's Medicare Claims Record Layout and Codebook

5.1.4.3 Physician Services

The carrier claims, known as the National Claims History (NCH) records, are largely from physicians although the file also includes claims from other non-institutional providers such as physician assistants, clinical social workers, nurse practitioners, independent clinical laboratories, ambulance providers, and stand-alone ambulatory surgical centers. Each carrier claim must include a Healthcare Common Procedure Coding System (HCPCS) code to describe the nature of the billed service. The HCPCS code is composed primarily of CPT-4 codes developed by the American Medical Association External Web Site Policy, with additional codes specific to CMS. Each HCPCS code on the carrier bill must be accompanied by an ICD-9 or ICD-10 diagnosis code (depending on the year; ICD-10 coding began in October 2015), providing a reason for the service. In addition, each bill has the fields for the dates of service, reimbursement amount, encrypted provider numbers (e.g., UPIN), and beneficiary demographic data. Note: in the most recent linkage, the UPIN has a new encryption scheme that is NOT compatible with previous linkage data.

Because of the large number of carrier claims, CMS maintains the data in variable length files. IMS, NCI's programming contractor, has converted these records into fixed length files by creating a record for each service that appears as a trailer on the CMS record. As a result, there may be multiple records for the same date of service. The variable `CLAIM_ID` was created to index unique claims. The variable `REC_COUNT` is a counter that enumerates each record associated with a claim, where `REC_COUNT` = 1 is also the first HCPCS in the first segment of a claim. The file is sorted by `PATIENT_ID`, `YEAR`, `CLAIM_ID`, and `REC_COUNT`.

View ResDAC's Carrier (Fee-For-Service) File Overview

View CCW's Medicare Claims Record Layout and Codebook

5.1.4.4 Home Health

The Home Health Agency file contains 100 percent of all claims for home health services. Some of the information contained in this file includes the number of visits, type of visit (skilled-nursing care, home health aides, physical therapy, speech therapy, occupational therapy, and medical social services), diagnosis (ICD-9 or ICD-10 diagnosis), the dates of visits, reimbursement amount, HHA provider number, and beneficiary demographic information. An HHA bill may cover services provided over a period of time, not a single day.

There are multiple parts to this file: base file, revenue center file, condition code file, occurrence code file, span code file, value code file, and demonstration/innovation code file.

View ResDAC's HHA File Overview

View CCW's Medicare Claims Record Layout and Codebook

5.1.4.5 Hospice

The Hospice file contains claims data submitted by Hospice providers. Some of the information contained in this file includes the level of hospice care received (e.g., routine home care, inpatient respite care), terminal diagnosis (ICD-9 or ICD-10 diagnosis), the dates of service, reimbursement amount, Hospice provider number, and beneficiary demographic information.

There are multiple parts to this file: base file, revenue center file, condition code file, occurrence code file, span code file, value code file, and demonstration/innovation code file.

View ResDAC's Hospice File Overview

View CCW's Medicare Claims Record Layout and Codebook

5.1.4.6 Durable Medical Equipment

The Durable Medical Equipment (DME) contains final action claims data submitted to Durable Medical Equipment Regional Carriers (DMERCs). Some of the information contained in this file includes diagnosis (ICD-9 or ICD-10 diagnosis), services provided (HCPCS codes), dates of service, reimbursement amount, DME provider number, and beneficiary demographic information. Claims for DME services that are processed by a carrier will be found in the NCH file. Claims for DME services that are processed by DMERCs will be found in the DME file. For example, claims for oral equivalents of IV chemotherapies will be found in the DME file.

There are multiple parts to this file: base file, line file, and demonstration/innovation code file.

View ResDAC's DME File overview

View CCW's Medicare Claims Record Layout and Codebook

5.1.4.7 Prescription Drug Events

Since July 2006, when Medicare coverage was expanded to include prescription drugs under Medicare Part D, approximately 60% of Medicare beneficiaries have enrolled in Part D. They either pay the Part D premium out-of-pocket or their premiums are paid for them, such as for low-income persons (i.e., dual enrollees). The Part D data included in SEER-Medicare begins in 2007.

Several files must be linked to analyze prescription drug details:

- Part D Drug Event File (PDE) Documentation - This file includes all transactions covered by Medicare prescription drug plan for both Prescription Drug Plans (PDPs) and Medicare Advantage Prescription Drug Plans (MA-PDs).
- Drug Characteristic File - variables appended to the PDE that describe the drug listed (e.g., NDC, brand and generic name)
- Formulary File Documentation - suite of three subfiles: formulary, excluded drug and Over the Counter Drug that contain information on how the plan covers the prescription drugs filled.
- Pharmacy Characteristics File Documentation - contains information about the pharmacy identified as the source of the drug for each PDE prescription fill record.
- Pharmacy Bridge File - In 2014, CMS changed the pharmacy identifier included on the PDE changed, this file provides a crosswalk that allows tracking the same pharmacy across this transition year.

  NOTE: Although one can track the same pharmacy over time, all pharmacy identifiers are encrypted.
- Prescriber Characteristics File Documentation - contains descriptive information for the prescriber identified in the PDE file.

View CCW's Medicare Claims Record Layout and Codebook

5.1.5 Medicare Assessment Data

Two sources of assessment data are provided in the most recent SEER-CAHPS linkage: the Minimum Dataset (MDS), a comprehensive, standardized assessment of nursing home residents' functional capabilities and health needs; and the Home Health Outcome and Assessment Information Set (OASIS),

which has information about patients' sociodemographic characteristics; health services utilization; physical and mental health; physical and cognitive function; comorbidities; physical, psychological, and psychosocial functioning; and living arrangements.

Individuals in the SEER-CAHPS data diagnosed with cancer in 1999 and later have been linked with MDS and OASIS data from 1999 and later. Both FFS and MA enrollees are included. MDS and OASIS data are also available from 1999 and later for persons included in the non-cancer sample.

View ResDAC's Long Term Care Minimum Data Set (MDS) 3.0 File Overview

View CCW's Assessment: MDS Record Layouts and Codebooks

View ResDAC's Home Health Outcome and Assessment Information Set (OASIS) File Overview

5.1.6 Information about Physicians, Hospitals, and Plans

A great deal of information is available on physicians and hospitals caring for Medicare beneficiaries, including the mergers & acquisitions file to track unique providers over time, and the hospital file to identify characteristics of hospitals (e.g., teaching status, size, Center of Excellence, etc.). All physician identifiers on the SEER-Medicare data are encrypted in order to protect the privacy of the physicians. Full details are available on the SEER-Medicare site: https://healthcaredelivery.cancer.gov/seermedicare/aboutdata/provider.html

Information about health plans is available in the SEER-CAHPS data; however, it is not possible to identify the names of plans or companies associated. Instead, plan identifiers are masked, unique strings. The following variables provide unique plan identifiers:

- SAMPLE: Plan ID of Surveyed Plan (`SA_PLAN_ID`)
- SAMPLE: Health Plan Contract Number (`SA_CONTRACT`)
- SAMPLE: Part A & B Contract Number (`SA_CONTRACT_AB`)
- SAMPLE: Part D Contract Number (`SA_CONTRACT_D`)

5.1.7 Area-level Characteristics: Geographic Data

Separate files are available that contain geographically-based (ZIP code and census tract level) socioeconomic information from the 1990 and 2000 Censuses and the 2008–2012 American Community Survey. These measures can be linked to individuals. Additional information is provided at: https://healthcaredelivery.cancer.gov/seermedicare/aboutdata/geographic.html

ZIP Code Census File Documentation (PDF, 51 KB)

Census Tract File Documentation (PDF, 52 KB)

6. How to Use SEER-CAHPS Data

This section provides a high-level overview of how to use SEER-CAHPS data.

6.1 Setting Up the Data for Analysis

SAS input and format statements are available here: https://healthcaredelivery.cancer.gov/seer-cahps/support/.

Please refer to **Section 6.6, Survey Analysis: Weights, Strata, and Methods** for how to set up your data as survey data and specify the appropriate weights and strata.

As discussed in **Section 4.2, Patient_ID: The Unique Identifier** , the `PATIENT_ID` variable allows analysts to link unique beneficiaries across all the files pertaining to beneficiaries. Merging various files 1:1 on `PATIENT_ID` is generally straightforward.

Please refer to **Section 6.8, Claims Analysis**for information specific to analyses of SEER-CAHPS FFS claims data, including deduplication and linking of claims files across time.

Please refer to **Section 6.5, Missing Data** for guidance on handling missing data.

The next sections provide broad guidance on using care experience measures in your analysis, covariate adjustment, and statistical modeling. Additional support is available via the online form: https://healthcaredelivery.cancer.gov/seer-cahps/contact.html or at the following email address: NCISEERCAHPS@mail.nih.gov.

6.2 Using Care Experience Measures in Your Analysis

Care experience measures are survey-reported measures of healthcare quality. They cover important aspects of high-quality care such as being able to get care when you need it, in a timely fashion, from a physician who communicates in a way you can understand. They are increasingly used in public reporting efforts and value-based care models. Care experiences differ from so-called "patient satisfaction" measures, as shown in **Table 7.**

Table 7. Differences between patient satisfaction and care experiences

Patient Satisfaction	Patient Experiences
Level of contentment with healthcare	Patient ratings of specific aspects of care
Whether a patient's expectations about a health encounter were met	Includes care from health plans, doctors, nurses, staff, healthcare facilities
Two people who receive same care but have different expectations may give different satisfaction ratings	Goal is to provide care that is respectful and responsive to individual patient preferences
"Did your doctor spend enough time with you?"	"When you needed care right away, how often did you get care as soon as you thought you needed?"

6.2.1 Overall Ratings

Overall ratings of care are summary measures meant to capture a respondent's overall sentiment. The wording of these items has changed over time; refer to the Details for Researchers file for year-by-survey details. Depending on year, plan, and survey, 5 overall ratings are available in SEER-CAHPS:

1. Using any number from 0 to 10, where 0 is the worst health care possible and 10 is the best health care possible, what number would you use to rate all your health care in the last 6 months? (RATE_CARE)
2. Using any number from 0 to 10, where 0 is the worst personal doctor possible and 10 is the best personal doctor possible, what number would you use to rate your personal doctor? (RATE_MD)
3. Using any number from 0 to 10, where 0 is the worst specialist possible and 10 is the best specialist possible, what number would you use to rate that specialist? (RATE_SPEC)
4. Using any number from 0 to 10, where 0 is the worst health plan possible and 10 is the best health plan possible, what number would you use to rate Medicare/health plan? (RATE_PLAN)
5. Using any number from 0 to 10, where 0 is the worst prescription drug plan possible and 10 is the best prescription drug plan possible, what number would you use to rate your prescription drug plan? (RATE_PDP)

6.2.2 Composite Scores

Individual Items and Related Items

Composite measures are calculated composites of individual items that are meant to capture a respondent's sentiment about a specific domain, or area of care. The wording of items has changed over time; refer to the Details for Researchers file for year-by-survey details. Depending on year, plan, and survey, 7 composite scores are available in SEER-CAHPS:

1. Doctor Communication (CMP_DRCOMM)
2. Getting Care Quickly (CMP_GETCAREQCK)
3. Getting Needed Care (CMP_GETNDCARE)
4. Getting Needed Prescription Drugs (CMP_GETNDDRG)
5. Care Coordination (CMP_CARECOORD)
6. Health Plan Information and Customer Service (CMP_CSTSRV)
7. PDP Information and Customer Service (CMP_PDCSTSRV)

How Composite Scores are Calculated

Composite scores are created using linear mean scoring. Details on which items are included in each measure and how each score is calculated are provided in the Details for Researchers file.

For SEER-CAHPS analyses, linear mean scoring is the preferred Medicare CAHPS scoring method. Please see https://healthcaredelivery.cancer.gov/seer-cahps/researchers/approaches_guidance.html for additional details.

Rescaling

The overall ratings and composite scores are on different scales: 0 to 10 (global) and 0 to 100 (composites; technically, 0-1, since these are percentiles). Some analysts choose to rescale one set or the other in order to put all the measures on the same scale.

Psychometricians have noted that the intervals between 0 and 5 vs. 6 and 10 are not necessarily the same, and that individuals vary in their tendency to report top-box (9 or 10) values. The measures are also highly skewed and non-normally distributed. For example, 6 of the care experience measures have may have median values equal to the maximum values (depending on your sample): overall ratings of personal doctor and specialist, and the composite measures for getting needed care, getting needed prescription drugs, doctor communication, and PDP customer service.

Caution is warranted when rescaling. Consider creating standardized scores (z-scores) or employing non-linear models to analyze these measures. See https://healthcaredelivery.cancer.gov/seer-cahps/researchers/approaches_guidance.html for additional details.

6.2.3 Reporting Data by Individual Cancer Site vs. Combined Sites

It is important to consider your research aims and questions to determine whether it is feasible and appropriate to include a cohort with multiple cancer sites versus a single cancer site. For example, different types of cancer (or different stages of the same types of cancer) may involve very different types of treatments; associated symptoms, morbidity, and mortality; and impacts on patient experience of care. These and other factors affect how clinically reasonable it is to combine data from patients diagnosed with different types (or stages) of cancer. For all analyses, you must also consider sample size and the purpose of your study.

Most SEER-CAHPS respondents are over age 65. Certain cancers are sex-specific (e.g., prostate and uterine cancers). These factors limit sample size. A power analysis can be helpful in determining minimum sample sizes. Consider in your power analysis that if your outcome measure is something that has little variation (e.g., certain care experience measures), you will need a much larger sample size to detect an effect. The online sample size estimator can provide an idea of how large a sample SEER-CAHPS may have for your particular research question.

6.3 Covariate Adjustment

6.3.1 Guidance on Standard Case-Mix Adjustment for SEER-CAHPS Analyses

Evaluations of patient experience surveys, including Medicare CAHPS, have identified respondent characteristics not under control of the health or drug plan but consistently related to the sampled member's survey responses, even among beneficiaries in the same health plan. Such associations may occur for a number of reasons:

- Beneficiaries with some characteristics may be more likely to encounter problems in health care (e.g., people requiring frequent care for chronic conditions)
- Beneficiaries with some characteristics may be treated differently than others (e.g., people who speak English as a second language)
- Some characteristics are associated with differences in the use of response scales (i.e., differential item response)[5]

Public reports of Medicare Advantage (MA) and Prescription Drug Plan (PDP) CAHPS Survey results are adjusted for the known effects of such characteristics. This process of case-mix adjustment helps to control for variability in patient experience ratings due to different distributions of patient characteristics known to be associated with patient experience scores. More information on case-mix adjustment as performed regularly by the MA & PDP CAHPS Project Team can be found on the Medicare Advantage and Prescription Drug Plan CAHPS® Survey site.

Case-mix adjustment variables for MA and PDP CAHPS Survey results

- Age
- Education
- General Health Status
- Mental Health Status

Case-Mix Adjustment

The standard case-mix variables have been shown to predict individuals' reports on their health care experiences and are generally acknowledged in the literature on patient experience reporting to represent characteristics largely defined prior to the period of care reported on. Inclusion of these covariates is appropriate for most analyses in which CAHPS measures are the outcomes (dependent variables), including multivariable analyses that use SEER-CAHPS data (*i.e.*, studies that use MA, PDP, and Fee-for-Service CAHPS data). We suggest using these set of covariates as a default setting for all analyses and always when making comparisons among health plans or other health care units.

- Received Help Responding
- Proxy Answered Questions for Respondent
- Medicaid (dual) enrollment
- Low Income Subsidy
- Chinese Language

6.3.1.1 Considerations

Case-mix adjustment variables added over time can be found on the Medicare Advantage and Prescription Drug Plan CAHPS® Survey site. When pooling data over multiple years, investigators are encouraged to include the covariates common to all years, which will often be the covariates used in the first year of CAHPS survey data requested. For example, if requesting CAHPS survey data from 2005-2013, investigators should use case-mix adjustment variables recommended in 2005. There may be certain situations where researchers choose to combine case-mix adjustment variables for analysis (e.g., dual enrollment and low-income subsidy; received help responding and proxy response status variables).

Analyses conducted using SEER-CAHPS data will typically involve cancer-related population subgroups and/or cancer-specific variables not used in standard CAHPS reporting analyses, coefficients of the estimates regressing CAHPS ratings and composites on standard case-mix variables will generally differ from those published in relation to the MA and PDP reports for the corresponding years.

Investigators should also consider additional covariates (e.g., cancer-specific variables) for inclusion in analyses. It is important to assess, however, whether these additional covariates are collinear with the standard case-mix adjustment variables. In addition, investigators are cautioned that inferences and interpretation of unadjusted CAHPS results are not appropriate.

6.4 Linear, Logistic, and Other Models

As discussed in **Section 6.2, Using Care Experience Measures in Your Analysis** , the care experience measures—as well as many other potential outcome measures, such as expenditures or utilization—are often skewed and non-normally distributed. The approach to modelling should be guided by the outcome measure and research aims, as well as a conceptual framework. Researchers are strongly advised to involve a statistician on their team.

6.5 Missing Data

Because SEER-CAHPS links data from multiple sources, there are different types of missing data in each data source. Below, we provide information on intended and unintended missing data, with recommendations for handling each type.

A video tutorial on missing data is available at https://healthcaredelivery.cancer.gov/seer-cahps/researchers/handling-missing-data.html.

The guidance in this user guide is excerpted from https://healthcaredelivery.cancer.gov/seer-cahps/researchers/missing-data-guidance.pdf (PDF, 526 KB); consult the full guidance for additional details.

Important note: Consistent with Medicare CAHPS guidance, we recommend never imputing CAHPS items or composites. The suggestions below only apply to other variables.

6.5.1 Types of Missing Data

Missing data are often categorized based on their mechanism: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). **Table 8** below defines each mechanism and explains how to test for the mechanism of your missing data. These tests are recommended for each new analytic sample.

Table 8. Missing data mechanisms

	Definition	Example	How to Determine
Missing Completely at Random (MCAR)	The propensity for a data point to be missing is completely random.	A survey respondent flips a coin to decide whether to complete a course evaluation.	Little's MCAR test (may not be totally definitive)
Missing at Random (MAR)	The propensity for a data point to be missing is not related to the missing data, but it is conditional on another variable.	Male respondents are more likely to decline to complete surveys, but declining *does not* depend on their level of satisfaction.	Test for interactions between observed variables: No significant interactions = MAR; Significant interactions = MNAR
Missing Not at Random (MNAR)	The propensity for a data point to be missing is not random.	Respondents with disabilities are less likely to complete surveys.	

6.5.2 Missing Data in SEER-CAHPS

In the SEER and Medicare enrollment files, missing data are generally designated with a separate category (for example, unknown stage) or a period (".") with no information on why a value might be missing. However, there are low fractions of missing information (FMI <1%) overall, since most of the information comes from administrative records that, because they are used for payment and reimbursement, are largely complete (from Medicare's perspective). Nevertheless, we only observe care that Medicare paid for.

In CAHPS, missing data on survey items are designated with a dot that is *sometimes* followed by a letter that provides additional information on why data are missing. It is possible to separate these types of missing data into intended and unintended types:

- Intended missing data occurs when the question was not on the survey, or the respondent had a valid skip or a valid answer of "don't know".
    - We recommend that analysts *not* impute these values.
- Unintended missing data arises when a respondent should have some data but does not, whether because they skipped it, refused, or gave an invalid response.
    - We recommend that analysts include such response values in a separate missing/unknown analytic category if the unintended FMI ≥ 25%.
    - If the unintended FMI < 25%, we recommend that analysts apply multiple imputation to that variable if the predictive imputation model appears to have validity.

Note that the MCAR, MAR, and MNAR categories are separate from intended/unintended. However, intended missing data are often MAR – for example, missingness is conditional on a variable such as survey year or type, but missingness is unrelated to care experiences.

Conversely, unintended missing data are often MNAR. For example, proxy respondents may skip or answer "don't know" to certain items AND proxy respondents generally perceive care quality as lower than do patient respondents. It is important to note that if MNAR data are handled as if they are MAR or MCAR, analysts are likely to arrive at inaccurate parameter estimates.

Table 9 below lists each type of missing data in CAHPS along with recommendations on how to handle missing values for each. Examples and further details are available in the full guidance document at https://healthcaredelivery.cancer.gov/seer-cahps/researchers/missing-data-guidance.pdf (PDF, 526 KB).

Table 9. Types of missing data in CAHPS and suggested methods for analysis

Missing value	Intended or Unintended Missing?	Suggested Analysis Method
. = Question Not on Survey	Intended	Do not impute. Exclude from denominator and "missing/unknown" category
.G = Good Skip based on Skip Pattern	Intended	
.V = Valid Answer of 'Does not apply'	Intended	
.D = Don't Know	Intended	Do not impute. OK to include in separate missing/unknown category
.N = Not Answered, on Survey	Unintended	Include in separate missing category if invalid FMI ≥ 25%; impute if invalid FMI < 25%
.R = Refused	Unintended	
.A = Answered-Should have Skipped	Unintended	
.S = Skipped-Should have Answered	Unintended	
.I = Inconsistent Response (to previous questions)	Unintended	
.O = Out of Range (Invalid value coded)	Unintended	
.M = Multiple Response	Unintended	
.Z = Provider Doesn't Match Survey Type	Unintended	

The full guidance also provides information on FMIs for 46 variables in the 2007-2013 data. Of those, 14 variables had 10% unintended missing or greater. The variable with the highest percentage of unintended missing data is self-reported cancer history (21%).

6.5.3 Considerations for Imputation

The main goals of the strategies for handling missing data are to minimize bias, maximize use of available information, and generate appropriate estimates of uncertainty (such as standard errors or confidence intervals). Many books and articles have been written about imputation. Common approaches to dealing with missing data include:

Complete case analysis (also known as listwise deletion)

- **Approach**: Drop cases with missing data on any variable of interest (done automatically in most software packages)
- **Drawbacks:** loss of data/observations; biased estimates unless data are MCAR

Unconditional mean imputation

- **Approach**: Replace missing values for a variable with its overall estimated mean
- **Drawbacks:** Artificially reduces variability; changes correlations between variables; may affect *P*-values and standard errors

Singular regression-based imputation

- **Approach**: Replace missing values with predicted scores from a regression equation
- **Drawbacks:** Decreases variability; underestimates uncertainty; may have dubious face validity if regression model does not fit data well (e.g., if the R^2 is low); inflates correlation between variables and biases R^2 statistics from analysis of imputed data

Stochastic imputation

- **Approach**: Add randomly drawn residual to imputed value from regression imputation. Distribution of residuals based on residual variance from regression model.
- **Drawbacks:** Standard errors are still attenuated (biased downward)

Multiple imputation

- **Approach**: Multiple values are imputed rather than a single value to reflect the uncertainty around the "true" value. Each imputed value includes a random component whose magnitude reflects the

extent to which other variables in the model cannot predict its "true" value. Variants include multiple imputation with chained equations (MICE) and Fully Conditional specifications that do not assume normal distributions for all variables and allow for different types of regression (linear, logistic, etc.) for imputation.

- **Drawbacks:** Auxiliary variables need to be correlated with missing variable (rule of thumb: r ≥ 40%). Biased estimates may result when N is relatively small and the FMI is high. Requires substantial computing power for larger Ns. Assumes data are MAR.

Newer methods of imputation are gaining proponents. Among them is multiple imputation using various machine learning methods, such as random forests (RF).6 Some researchers have found that RF imputation produces less biased results with narrower confidence intervals than regression-based imputation.7 Evidence suggests that RF-based imputation methods may be theoretically sound even for large percentages of missing values (up to 50%).6,8

6.5.4 Missing Data: Final Thoughts

In the SEER-CAHPS 2007-2013 sample, a little less than a third of major predictor variables had more than 10% invalid missing data, and none had more than 21%. However, when combining both intended and unintended missingness types, up to 96% of respondents have missing data; some variables, such as limitations in social activities, may have particularly high total FMIs across pooled-year samples because they were asked in only one year.

One question that is often raised by reviewers is how much data are missing from particular covariates. We would advise that analysts using the SEER-CAHPS data distinguish between intended and unintended missing when tabulating missingness in their articles for publication. This may pre-emptively address concerns about missing data that are endemic to survey research.

Distinguishing between intended and unintended missing data is challenging but important in any analysis. It is particularly important when using methods that impute missing data by default. Analysts using the SEER-CAHPS data resource would be advised to decide in advance whether to use imputation and how to account for missing data on key predictors.

6.6 Survey Analysis: Weights, Strata, and Methods

One of the features of the SEER-CAHPS data resources is the availability of data on Medicare FFS and MA enrollees with and without Prescription Drug Plans (PDPs). However, each MA plan has to survey a representative sample of its insurees, so the MA population is over-sampled relative to those with FFS Medicare. In order to produce estimates that better represent the distribution of FFS and MA enrollees in the Medicare population, the SEER-CAHPS data provides two different weight variables.

WGT_SIMPLE is a base weight calculated to make the sample representative of the beneficiary populations in the units in the original design. All years and survey types have this type of weight.

Using the base weight variable allows analyses to produce estimates that are representative of the beneficiary populations in the units of the original design. For the MA and standalone PDP sample, these units were contracts; for the FFS sample, these were states.

WGT_RAKED was constructed after using a raking weighting procedure (loglinear weights calculated by iterative proportional fitting) to weight the respondents to match the control distributions estimated from the first-round sample (with base weights). In some cases, small cells were collapsed with adjacent cells to avoid extreme weights. MA and FFS 2000-2004 do not have this type of weight as the group calculating the weights was unable to get data on non-respondents from that period. Using the raked weight variable allows analyses to correct for biases arising from differential nonresponse associated with beneficiary characteristics and reduces the effects of random variation in nonresponse. Currently, raked weights are only available for respondents with surveys in 2007 or later.

Both sets of SEER-CAHPS weights described above have been calibrated to the survey populations and sub-populations. The weights take nonresponse and strata characteristics into account. The calculation algorithm ensures that variance estimates for survey responses *within a subset of the data* are preserved* regardless of the size and characteristics of the dataset to which that subset belongs. Thus, no further calculation of survey weights is necessary. Further, the primary sampling units (PSUs) and strata that correspond to those weights are included in the SEER-CAHPS data linkage; no additional information regarding population or subpopulation size is required. The variables used to specify PSUs, strata, and weights are listed below.

*Note: small differences in variance estimates *may* be observed in calculations performed using different software platforms (*e.g.*, SAS *vs.* SUDAAN *vs.* STATA *vs.* R). Those differences are typically too small to meaningfully impact analytic results. However, if absolute consistency across statistical software packages is desired, it is recommended that researchers explore calculation-algorithm options within the packages being used, as default options for calculating variance estimates differ across platforms. Specifying options to be consistent across platforms may resolve those analytically minor differences in variance estimates.

To specify the sample design when analyzing the data, the following variables are suggested:

- Primary Sampling Unit: `PATIENT_ID`
- Strata
  - FFS without a PDP: `SA_FIPS_STATE`
  - FFS with PDP or standalone PDP: `SA_CONTRACT`
  - MA: `SA_PLAN_ID`

- Weights
  - Surveys from 2007 and later: `WGT_RAKED`
  - Surveys before 2007: `WGT_SIMPLE`

Additional information can be found in the yearly Medicare CAHPS reports related to Weighting.

 MA & PDP CAHPS Individual-Level Weight Construction (PDF)

The following text may be used in describing the weights briefly in manuscripts:

> "Data were weighted to represent the enrolled population of state (for FFS) or contract (for MA and PDP). For respondents in 2011 and later, weights were generated by applying a raking procedure (loglinear weights by iterative proportional fitting) to respondents to match weighted sample distributions within each contract (or state, for FFS beneficiaries) of gender, age, race/ethnicity, Medicaid and low income supplement eligibility, Special Needs Plan status, PD enrollment, and zip-code level distributions of income, education, and race/ethnicity."

6.7 Small Sample Size Cell Suppression

In order to protect the confidentiality and identity of patients, health plans, and providers, cell sizes of less than 11 in a table must be suppressed in accordance with the SEER-CAHPS Data Use Agreement. No cell containing a value of 1 to 10 can be reported directly. In addition, no cell can be reported that allows a value of 1 to 10 to be derived from other reported cells or information (i.e., use of percentages or other mathematical formulas that would allow the derivation of patient, facility or provider counts of less than 11). The cell suppression policy also applies to the reporting of excluded cases. There are several options that can be used to comply with these requirements, including collapsing cells, coarsening data, and cell suppression; the CMS Research Data Assistance Center (ResDAC) has more information.

6.8 Claims Analysis

Medicare claims data provide a rich, detailed portrait of beneficiaries' healthcare utilization. **Section 5.1.4, Medicare Claims Data** describes each of the claims data files in SEER-CAHPS. In this section, we briefly summarize aspects of claims analysis that are specific to the SEER-CAHPS data.

Not every SEER-CAHPS analysis will require claims. The ones that do look at claims should aim to understand aspects of care that cannot be adequately captured any other way. For example, researchers interested in surgical utilization can potentially observe cancer-directed surgery in the SEER data. However, any aspect of chemotherapy will likely need to analyze claims data, since SEER data do not capture chemotherapy. Another example would be hospitalization: the CAHPS FFS-only survey included a self-reported item in 2000-2004 on whether the respondent had any overnight hospital stay in the past 12 months (`INPAT`). Any analysis of inpatient stays among FFS beneficiaries outside of that small group of respondents will require a claims analysis.

6.8.1 Continuous Enrollment

If you analyze FFS claims, you will need to check whether beneficiaries had continuous coverage in Parts A (inpatient) and B (outpatient) as well as no HMO indicators during any part of your claims period. Otherwise, you may miss some care provided when an enrollee was not in FFS, and thus does not have claims data available. These indicators are provided in the MBSF for each year of enrollment and are as follows (with short variable names):

- `A_MO_CNT`: the number of months during the year that the beneficiary had Medicare Part A coverage
- `B_MO_CNT`: number of months during the year that the beneficiary had Medicare Part B coverage
- `HMO_MO`: the number of months during the year that the beneficiary received their Part A and Part B benefits through a managed care plan (i.e., a Medicare Advantage [MA] plan) instead of the traditional fee-for-service (FFS) program. Any month where the HMO indicator variable (HMO_IND_XX) was anything other than a 0 (not a member of an HMO) or a 4 (FFS participant in a case or disease management demonstration project) is counted as a MA month.

6.9 Health Status and Conditions

Most analyses of cancer populations seek to understand, and adjust for, health status, comorbidity, and activity limitations that may affect outcomes. A substantial body of literature has shown the importance of multimorbidity to every phase of the cancer care continuum – from screening, to diagnosis, to treatment, to survival. Numerous individual items describing health status are available in SEER-CAHPS, with varying degrees of completeness.

Do the SEER-CAHPS data identify comorbidities that are present before and after cancer diagnosis?

Although the data vary from year to year, some years of SEER-CAHPS data do contain indicators for a limited set of self-reported comorbidities, including heart attack, angina, COPD, and diabetes. The question asked was: "Has a doctor ever told you that you had any of the following conditions:…" The CAHPS surveys also elicit other self-reported health-related information, such as general and mental health status; limitations in activities of daily living, such as bathing, dressing, eating, and toileting; presence of any chronic condition and related doctors' visits; and smoking status. The questions asked varied from year to year and survey version to survey version. They may be identified before or after cancer diagnosis, depending upon the timing of the CAHPS survey. Please refer to the CAHPS Details for Researchers file for complete details on the data available for your study period.

For FFS enrollees, researchers can calculate any claims-based measure of morbidity, such as the NCI-Combined Index, the Charlson comorbidity index, CMS-HCC scores, or other scores produced by risk-

adjustment software. See SEER-Medicare: Calculation of Comorbidity Weights for more information. Although validated and commonly used, these comorbidity indices require claims data. Researchers who depend on such measures of morbidity will, thus, be limited to analyses of FFS enrollees with complete claims available during the comorbidity measurement period.

For the approximately 60% of the sample that is enrolled in Medicare Advantage (and thus do not have any linked claims or encounter data available), the morbidity and utilization data available for most individuals are limited to the self-reported CAHPS survey data and SEER-collected clinical data about each beneficiary's cancer (some beneficiaries also have MDS and OASIS data).

As an alternative to claims-based measures, the SEER-CAHPS team has developed the SEER-CAHPS Illness Burden Index (SCIBI).[9] The SCIBI is a machine-learning-derived summary score that approximates relative risk of mortality within 12 months after survey response. The SCIBI allows researchers to analyze illness burden information for MA enrollees as well as FFS enrollees. Future efforts are planned to update the SCIBI to include indicators from the OASIS and MDS data as well.

6.9.1 The SEER-CAHPS Illness Burden Index (SCIBI)

SEER-CAHPS Illness Burden Index (SCIBI) scores are currently available for individuals surveyed in 2007 and later. They incorporate whatever information is available for a respondent in terms of self-reported and claims information, including activities of daily living (ADL) limitations, other limitations in activities, self-reported conditions, and healthcare utilization (both claims-based and self-reported).

Versions

Two versions are available to users requesting the SEER-CAHPS data:

- Concurrent Basic (SCIBI-CB): These scores include predictor data from the 12 months before ***and after*** the survey response in the predictions; such indicators as hospice and DME use are measured across the full 24-month period
- Prospective Basic (SCIBI-PB): These scores only include predictor data from the 12 months before survey response

Normalized Z-scores

The SCIBI scores vary in their distributions depending on year, whether a person is in MA or FFS, and whether they were surveyed before or after their cancer diagnosis. Thus, in addition to the cohort-specific raw scores (developed within each year-group slice), we provide normalized z-scores centered on the population mean (i.e., people with and without cancer, MA and FFS, all years). These z-scores, which have a mean of 0 and an SD of 1, can thus be used to compare illness burden using the same “measuring stick” for every person in the linked data.

SCIBI in Action

Please see the 2019 methods paper for an overview of how the SCIBI was developed.[9]

The SCIBI has been used in an analysis of SEER-CAHPS using data from 2007-2015 (manuscript in press; abstract presented at the American Public Health Association's 2020 annual meeting).[10] The research team found that, after controlling for other case-mix adjustors, higher SCIBI scores (indicating greater illness burdens) were significantly associated with better ratings of Health Plan and better Getting Care Quickly scores. In Bayesian models, individuals with higher illness burden had similar results on the same two measures and also reported reliably worse Overall Care experiences. These results suggest that illness burden may influence how people experience care or report those experiences, independently of standard case-mix adjustors (such as self-reported general and mental health status).

6.10 Proxy Responses

SEER-CAHPS survey respondents may be unable to complete surveys without assistance for one or more reasons, including but not limited to lack of or limited English proficiency, difficulty with reading or writing, and acute or chronic medical conditions that impair the ability to respond. In such cases, they may have a designated *proxy* to assist in some or all of the tasks required to respond to the SEER-CAHPS survey.

Survey responses provided by a proxy or with the help of a proxy provide a source of information about patient characteristics and care experiences for individuals that might otherwise be unavailable. However, caution should be taken for respondents for whom proxy use is indicated, as proxy responses may vary from self-report in systematic ways.

Proxy ratings of care have been shown to be significantly less positive evaluations of care experiences relative to self-report.[11] Lines and colleagues found that proxy use was significantly more prevalent among dual-Medicare/Medicaid eligible SEER-CAHPS respondents relative to non-dual enrollees;[4] in a separate study, Lines and colleagues also showed that proxy use *itself* was an important predictor of illness burden.[12] In a study of Medicare beneficiaries, levels of agreement between proxy responses and self-report depended on the content elicited by specific questions: proxy responses on survey items regarding sensory status of the respondent, simple questions, and questions about observable phenomena were more reliably consistent with self-report than questions regarding cognitive, physical, or social status; complex questions; or questions eliciting personal or private information.[13] Thus, researchers are advised to be cautious with data elements given by proxy or with the aid of a proxy.

Proxy status is indicated in the SEER-CAHPS dataset by an overall measure and a series of variables describing proxy activities. The variable PROXY is a binary indicator of whether a proxy helped the respondent in any way, based on the individual proxy items listed below. A value of "0" for PROXY indicates that the respondent did not make any use of a proxy; when this is the case, the fields for the other proxy variables will be empty. A value of "1" for PROXY indicates that a proxy helped the respondent; when this is

the case, the remaining proxy variables will indicate the type of help provided by the proxy. PROXY is missing in some cases; refer to **Section 6.5, Missing Data** for more details on handling missing data.

Refer to the Details for Researchers file for specific item availability by year, survey/population, and changes in wording over time.

The individual proxy items in the SEER-CAHPS dataset include:

- Variable indicating that a proxy answered questions on behalf of the respondent (PXY_ANSW: 1 = marked, 0 = not marked)
- Variable indicating the type of help proxy provided (PXY_HELP: 1 = Read/Wrote, 2 = Answer Questions, 3 = Other way)
- Variable indicating that a proxy helped in a way other than the options listed (PXY_OTHR: 1 = marked, 0 = not marked)
- If PXY_OTHER = 1, a variable specifying the other way(s) a proxy helped the respondent (PXY_OTHR_TXT)
    - <Missings>
    - '1' = 'Read the questions to me'
    - '2' = 'Wrote down the answers I gave'
    - '3' = 'Answered the questions for me'
    - '4' = 'Translated the questions into my lang'
    - '5' = 'Helped in some other way'
    - '6' = 'Helped number'
    - '7' = 'Discussed/Explained'
    - '8' = 'None/Not Applicable'
    - 'A' - 'zzz' = 'Other Specify'
- Variable indicating that a proxy read questions for the respondent (PXY_READ: 1 = marked, 0 = not marked)
- Variable indicating that a proxy wrote answers for the respondent (PXY_WRIT: 1 = marked, 0 = not marked)
- Variable indicating the proxy helper relationship to the respondent (PXY_RELATION)
    - <Missings>
    - 1 = 'Spouse/life partner'
    - 2 = 'Parent'
    - 3 = 'Child'
    - 4 = 'Other family member'
    - 5 = 'Friend'
    - 6 = 'Roommate or housemate'
    - 7 = 'Employee'
    - 8 = 'Employer'

    - 9 = 'Health care worker'
    - 10 = 'Other'

- Item indicating that a proxy translated questions for the respondent (PXY_TRANS: 1 = marked, 0 = not marked)

7. References

1. Freed M, Koma W, Cubanski J, Fugelsten Biniek J, Neuman P. More Than Half of All People on Medicare Do Not Compare Their Coverage Options Annually. 2020; https://www.kff.org/medicare/issue-brief/more-than-half-of-all-people-on-medicare-do-not-compare-their-coverage-options-annually/. Accessed Apr. 7, 2021.
2. Cubanski J, Damico A. Medicare Part D: A First Look at Medicare Prescription Drug Plans in 2021. 2021; https://www.kff.org/medicare/issue-brief/medicare-part-d-a-first-look-at-medicare-prescription-drug-plans-in-2021/. Accessed Apr. 7, 2021.
3. Freed M, Damico A, Neuman P. A Dozen Facts About Medicare Advantage in 2020. 2021; https://www.kff.org/medicare/issue-brief/a-dozen-facts-about-medicare-advantage-in-2020/. Accessed Apr. 7, 2021.
4. Lines LM, Cohen J, Halpern MT, Smith AW, Kent EE. Care experiences among dually enrolled older adults with cancer: SEER-CAHPS, 2005–2013. *Cancer Causes & Control.* 2019;30(10):1137–1144
5. Elliott MN, Haviland AM, Kanouse DE, Hambarsoomian K, Hays RD. Adjusting for subgroup differences in extreme response tendency in ratings of health care: impact on disparity estimates. *Health services research.* 2009;44(2p1):542-561.
6. Tang F, Ishwaran H. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal.* 2017;10(6):363-377.
7. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology.* 2014;179(6):764-774.
8. Bodner TE. What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal.* 2008;15(4):651-675.
9. Lines LM, Cohen J, Halpern MT, Kent EE, Mollica M. Random Survival Forests Using Linked Data to Measure Illness Burden Among People With Cancer: Development and Internal Validation of the SEER-CAHPS Illness Burden Index. Oral presentation presented at AcademyHealth Annual Research Meeting; June 4, 2019, 2019; Washington, DC.
10. Lines L, Kirschner J, Halpern M, Kent EE, Mollica M, Smith AW. Illness burden is associated with care experience measures among Medicare beneficiaries surveyed before or after a cancer diagnosis. American Public Health Association Annual Meeting; 2020; Virtual.
11. Elliott MN, Beckett MK, Chong K, Hambarsoomians K, Hays RD. How Do Proxy Responses and Proxy-Assisted Responses Differ from What Medicare Beneficiaries Might Have Reported about Their Health Care? *Health Services Research.* 2008;43(3):833-848.
12. Lines LM, Cohen J, Kirschner J, et al. Random survival forests using linked data to measure illness burden among individuals before or after a cancer diagnosis: Development and internal validation of the SEER-CAHPS illness burden index. *International Journal of Medical Informatics.* 2021/01/01 2021;145:104305.
13. Li M, Harris I, Lu ZK. Differences in proxy-reported and patient-reported outcomes: assessing health and functional status among medicare beneficiaries. *BMC medical research methodology.* 2015;15(1):1-10.

About this User Guide

This user guide was written and produced by researchers from the National Cancer Institute's Healthcare Delivery Research Program, which oversees the SEER-CAHPS data resource (MTH, JM, MAM), and from RTI International, under contract (#75N91020F00215) to NCI (DHB, KD, LD, RL, LML, DZ).

In alphabetical order, contributors included Daniel H. Barch, PhD; Kim Danforth, ScD, MPH; Lisa DiMartino, PhD, MPH; Michael T. Halpern, MD; Rebecca Lewis, MPH; Lisa M. Lines, PhD, MPH; Joshua Medel, MS; Michelle A. Mollica, PhD, MPH, MSN, OCN; and Diana Zabala, BS.

If you have any questions about this guide, suggestions, or error corrections, please contact NCISEERCAHPS@mail.nih.gov.

Version 1.2 Published: DD MM YYYY

DOI:10.17917/SNGP-FQ26

Suggested citation: National Cancer Institute. SEER-CAHPS User Guide v.1.0. 2021. Available at: https://healthcaredelivery.cancer.gov/seer-cahps/researchers/guidance.html. Accessed [date]. doi: 10.17917/SNGP-FQ26.