



# **Chronic Condition Data Warehouse**

*Your source for national CMS Medicare and Medicaid research data*

## **CCW Technical Guidance: Calculating Medicare Population Statistics**

January 2018

Version 1.2

This page intentionally left blank.

## Revision History

Revision Date	Version Number	Description
01/2018	1.2	MBSF A/B references updated to include C/D; updated data source information to reflect CME instead of EDB.

## TABLE OF CONTENTS

<b>Chapter 1: Overview</b> .....	<b>1</b>
<b>Chapter 2: CCW Medicare Data</b> .....	<b>5</b>
A. Master Beneficiary Summary File A/B/C/D Segment .....	6
B. Master Beneficiary Summary File CCW Chronic Condition Segment .....	7
C. Master Beneficiary Summary File Other Chronic or Potentially Disabling Conditions Segment .....	7
<b>Chapter 3: Calculating Population Statistics</b> .....	<b>9</b>
A. Measures .....	9
1. Numerators .....	9
2. Denominators .....	11
B. Analysis of Rates .....	12
1. Description of Options .....	12
2. Calculation of Options .....	14
<b>Chapter 4: Results</b> .....	<b>21</b>
<b>Chapter 5: Generalizing These Methods</b> .....	<b>23</b>
A. Limitations .....	24
<b>Chapter 6: Further Assistance with Medicare Administrative Claims Data</b> .....	<b>25</b>
<b>Appendix A: List of Acronyms and Abbreviations</b> .....	<b>26</b>

## LIST OF TABLES

Table 1. CCW Chronic Disease Classifications .....	2
Table 2. CMS’s Other Chronic or Potentially Disabling Conditions.....	2
Table 3. Four Methods for Calculating Prevalence of AMI in 2015 (5% random sample).....	21

## LIST OF CODE EXAMPLES

Code Example 1. Define Sample of Beneficiaries with Full or Partial Medicare Fee-For-Service (FFS) Coverage Using Variables from MBSF.....	16
Code Example 2. Define Sample of Beneficiaries with Medicare Fee-For-Service (FFS) Coverage at a Point in Time .....	18
Code Example 3. Determine Total Member Months of Medicare Fee-For-Service (FFS) Coverage, and Corresponding Condition Numerator .....	19

## Chapter 1: Overview

The Centers for Medicare & Medicaid Services (CMS) launched the Chronic Condition Data Warehouse (CCW), a research database, in response to the Medicare Modernization Act of 2003 (MMA). Section 723 of the MMA outlined a plan to improve the quality of care and reduce the cost of care for chronically ill Medicare beneficiaries. An essential component of this plan was to establish a data warehouse that contains Medicare claims data and assessments, linked by beneficiary, across the continuum of care.

The CCW contains fee-for-service (FFS) institutional and non-institutional claims, assessment data, and enrollment/eligibility information from January 1, 1999 forward. The CCW data are linked by a unique, unidentifiable beneficiary key, which allows researchers to analyze information across the continuum of care. CCW data are available upon request for a random 5% sample or for specific chronic condition cohorts. Data may also be requested for other cohort(s) of interest. The twenty-seven (27) predefined CCW common and chronic condition categories<sup>1</sup> were defined by CMS, in collaboration with the Research Data Assistance Center (ResDAC) and the CCW contractor<sup>2</sup>. Much of the supporting documentation is available on the CCW website (e.g., chronic condition definitions, data dictionary, etc.). See [www.ccwdata.org](http://www.ccwdata.org). CMS developed additional measures for Other Chronic or Potentially Disabling Conditions specifically to enhance research of the Medicare-Medicaid dually enrolled population. These additional conditions include variables for mental health and tobacco use conditions, developmental disorder and disability-related conditions, and other chronic physical and behavioral health conditions.

The CCW data files were designed to support a variety of research objectives. The pre-defined conditions make it easy to select a study population with a condition of interest. Researchers may refine the clinical or coverage criteria as needed for their specific hypotheses. Medicare claims-based utilization information is used to make the condition determinations (i.e., an indicator that the beneficiary received a service or treatment for the condition of interest).

The 27 CCW conditions are listed in Table 1. The algorithms for identification of these condition categories are available on the CCW website (<https://www.ccwdata.org/web/guest/condition-categories>).

---

<sup>1</sup> As a historical note, when CCW data first became available there were only 21 CCW conditions. When 2010 data became available, we revised 17 of the original 21 chronic conditions (CCs); the updated 27 CCs are available from 1999 forward.

<sup>2</sup> The original contractor for the CCW was the Iowa Foundation for Medical Care.

**Table 1.** CCW Chronic Disease Classifications

Acquired Hypothyroidism*	Chronic Kidney Disease
Acute Myocardial Infarction	Chronic Obstructive Pulmonary Disease & Bronchiectasis (COPD)
Alzheimer's Disease	Depression
Alzheimer's Disease, Related Disorders, or Senile Dementia	Diabetes
Anemia*	Glaucoma
Asthma*	Heart Failure
Atrial Fibrillation	Hip / Pelvic Fracture
Benign Prostatic Hyperplasia*	Hyperlipidemia*
Cancer, Colorectal	Hypertension*
Cancer, Endometrial	Ischemic Heart Disease
Cancer, Female/Male Breast	Osteoporosis
Cancer, Lung	Rheumatoid Arthritis / Osteoarthritis
Cancer, Prostate	Stroke / Transient Ischemic Attack
Cataract	

\* The asterisk indicates the condition was not one of the original 21 conditions.

The CMS Other Chronic or Potentially Disabling conditions are listed in Table 2. The algorithms for identification of these condition categories are also available on the CCW website (<https://www.ccwdata.org/web/guest/condition-categories>).

**Table 2.** CMS's Other Chronic or Potentially Disabling Conditions

ADHD, conduct disorders, and hyperkinetic syndrome	Mobility impairments
Anxiety disorders	Multiple sclerosis and transverse myelitis
Autism spectrum disorders	Muscular dystrophy
Bipolar disorder	Obesity
Cerebral palsy	Other developmental delays
Cystic fibrosis and other metabolic developmental disorders	Peripheral vascular disease
Depressive disorders	Personality disorders
Epilepsy	Post-traumatic stress disorder (PTSD)
Fibromyalgia, chronic pain and	Pressure ulcers and chronic ulcers
	Schizophrenia
	Schizophrenia and other psychotic

chronic fatigue	disorders
Hepatitis (any viral type)*	Sensory impairment (blindness and visual)
HIV/AIDS **	Sensory impairment (deafness and hearing)
Intellectual disabilities and related conditions	Spina bifida and other congenital anomalies of the nervous system
Learning disabilities and other developmental delays	Spinal cord injury
Leukemia and lymphoma	Tobacco use disorders
Liver disease, cirrhosis, and other liver conditions (excluding hepatitis)	Traumatic brain injury and nonpsychotic mental disorders due to brain damage
Migraine and other chronic headaches	

*\*In addition to the aggregate Hepatitis condition variable, the MAX and MMLEADS data files have fields for various hepatitis subtypes (e.g., Hepatitis A, Hepatitis B [acute], Hepatitis B [chronic]).*

*\*\*In addition to the HIV/AIDS condition variable, the MAX and MMLEADS data files have fields for a variety of HIV/AIDS screening and services - including several "monitoring" variables (largely consisting of Medicaid services).*

The variables for the CMS Other Chronic or Potentially Disabling Conditions are similar to those for the CCW conditions in that they are generated using claims-based algorithms which indicate that treatment for a condition appears to have taken place. The algorithms examine patterns of Medicare services, which serve as a proxy indicating that a beneficiary is likely receiving treatment for the condition<sup>3</sup>. You can find more information about these algorithms, including the literature references and exact codes and claim types used to identify each condition, on the CCW website (<https://www.ccwdata.org/web/guest/condition-categories>).

It is important to note that the major objective of the chronic condition indicator variables is to allow for relevant clinical cohorts to be easily extracted from a very large database. The CCW definitions were intended to be somewhat broad, so that more researchers could request data extractions based on these definitions – then refine the specifications as needed to fit their own data needs. These specifications are communicated using the CCW Specifications Worksheet that is part of the CMS data request packet. For this reason, the counts published in Medicare data tables and charts on the CCW website (see <https://www.ccwdata.org/web/guest/medicare-tables-reports>) for the various chronic conditions are likely overestimates of the *useable* sample for a given research project. Correspondingly, it would not be accurate to use these chronic

<sup>3</sup> A version of the CCW Conditions and CMS Other Chronic or Potentially Disabling Conditions that examines both Medicare and Medicaid data for those dually enrolled in Medicare and Medicaid is available in the MMLEADS data product.



condition counts for the purposes of calculating population statistics for these conditions without first narrowing the counts further by using various criteria.

If it is desirable to have a control group, there are some simple steps that can be taken to identify the appropriate controls. Researchers should consider a variety of potential control factors, such as the presence or absence of other conditions – and whether current treatment or “ever” treatment for a comorbid condition is of interest, the length of observation or surveillance periods, and certain types of Medicare coverage.

This paper is intended to provide guidance to researchers in order to:

- 1) appropriately identify a study group based on clinical criteria (or combination of clinical and coverage specifications), and
- 2) define an appropriate denominator for rate calculations, if applicable.

## Chapter 2: CCW Medicare Data

CCW contains data for 100% of Medicare beneficiaries, including enrollment data and FFS claims from 1999 forward. CCW has been designed to extract standard samples of beneficiaries, such as the CMS 5% random sample, an *enhanced* 5% sample (i.e., once a beneficiary is selected, he/she remains part of this sample in all subsequent years; once-in-always-in rule, from 1999 forward), or a user-defined cohort that can be extracted from the CCW 100% Medicare files. Data for all eligible beneficiaries are contained within the CCW database; it is not limited to just those with a chronic condition. Note that claims for services provided to Medicare beneficiaries with managed care plans are not included in the CCW; therefore the CCW should be viewed as a source of utilization information primarily for the fee-for-service (FFS) population. Assessment information is available for all who have received the prerequisite services, and not affected by this limitation.

The predefined chronic conditions use claims-based definitions; therefore, there is not an opportunity to determine whether Medicare managed care enrollees have the condition(s) of interest. This limitation also applies, perhaps to a lesser extent, to newly-eligible Medicare beneficiaries who may have only a partial year of FFS coverage.

The CCW extracts are provided to researchers in a user-friendly format. SAS<sup>®</sup> read-in statements are provided along with the data files requested by the researcher. Although two different versions of read-in statements are routinely provided – the SASv6 file which contains traditional short SAS<sup>®</sup> names, and the SASv8 read-in statements which take advantage of newer features in SAS<sup>®</sup> allowing for longer and more descriptive variable names, the variable names used for examples in this document are the long names found in the SASv8 read-in file.

The CCW has always disseminated files which include data regarding Medicare enrollment. Historical files include the CMS Denominator File, the CCW Beneficiary Summary File (BSF), the Beneficiary Annual Summary File (BASF), and the Master Beneficiary Summary File (MBSF) that used the CMS Enrollment Database (EDB) as its source. Currently, the preferred enrollment data file is the MBSF that uses the CMS Common Medicare Environment (CME) Database as its source. The CCW transition from the EDB to the CME database as the source for the MBSF enrollment and demographic information offers a number of advantages that will better meet researchers' evolving data needs. A description of the rationale and impact of this data conversion is contained in a CCW White Paper, [Medicare Enrollment: Impact of Conversion from EDB to CME](#).

Like the original MBSF, the current MBSF using the CME enrollment data file contains a constellation of enrollment and other person-level variables contained in “segments” – which are separate components of the file researchers may elect to purchase. A description of the variables contained in the MBSF can be found on the Data Dictionary page on the CCW website at <http://www.ccwdata.org/web/guest/data-dictionaries>.

The MBSF is created annually and contains demographic entitlement and enrollment data for beneficiaries who: 1) were part of the user-requested sample; 2) were documented as being alive for some part of the reference year; and, 3) were enrolled in the Medicare program during the file’s reference year. Reference year refers specifically to the calendar year accounted for in the MBSF. For example, the 2014 MBSF covers the year 2014 – which is the reference year.

The current MBSF – Part A/B/C/D segment, also known as the Base Beneficiary Summary File, consists of variables that identify monthly Medicare Part A, B, C, and D enrollment status and other key demographic and coverage variables. All of the fields regarding beneficiary demographics or enrollment in the MBSF are derived from the CMS Common Medicare Environment (CME) Database (which is also the source for the CMS Enrollment Database – EDB). CCW updates information each month, and allows a full year of additions and updates after the end of the calendar year before finalizing the MBSF. For example, the 2014 data were finalized in December 2015 and then extracted to populate the MBSF.

The additional segments of MBSF are: 1) CCW Conditions, 2) Other Chronic or Potentially Disabling Conditions, and 3) Cost and Use.

The Base Beneficiary Summary File (also called the Medicare Part A, B, C and D Enrollment Segment) consists of variables related to Medicare Part A, Part B and Part D enrollment data. The CCW Chronic Condition Segment and the Other Chronic or Potentially Disabling Conditions Segment of the MBSF contain variables which indicate the presence of treatment for common or chronic conditions using claims-based algorithms (as a proxy for evidence of the presence of a condition).

## **A. Master Beneficiary Summary File A/B/C/D Segment**

CCW creates the Base Master Beneficiary Summary File as a calendar year file. It contains demographic, geographic, eligibility and Medicare Part A, B, C and D coverage information for all beneficiaries who are alive and eligible for any part of the year.

This file contains a variable to ascertain whether a beneficiary was included in the CMS 1%, 5%, or 20% sample for the year (i.e., the cross-sectional annual sample; variable called

SAMPLE\_GROUP) or if the beneficiary was included as a member of the *enhanced* CCW 5% sample (i.e., all inclusive, *ever* included in the 5% sample, from 1999 forward; variable called ENHANCED\_FIVE\_PERCENT\_FLAG). Researchers may use this annual person-level summary file to determine whether a beneficiary has a sufficient surveillance period (i.e., months of Medicare coverage) for inclusion in the study. Variables contained in this file include the number of months of Medicare Part A, Part B, or managed care coverage, whether the beneficiary died during the year, as well as other beneficiary demographic and geographic information (see for the record layout).

## **B. Master Beneficiary Summary File CCW Chronic Condition Segment**

The CCW Chronic Condition Segment of the MBSF contains summarized clinical information for all beneficiaries included in the requested cohort. This file includes a set of three variables for each of the 27 chronic conditions. These three variables include: 1) a yearly indicator – which indicates whether each of the 27 chronic condition definitions was met during the respective time period ending December 31, YYYY, 2) a mid-year indicator – which may be useful if researchers are using a July 1 time frame, and 3) a first occurrence date – which indicates the date the beneficiary was first identified as having met the specifications for the condition (note: 1999 is the earliest year that will appear in this field). Please note that these chronic condition fields are defined by looking at a pattern of medical care utilization, as determined by Medicare FFS claims. See <https://www.ccwdata.org/web/guest/data-dictionaries> for the record layout.

## **C. Master Beneficiary Summary File Other Chronic or Potentially Disabling Conditions Segment**

The Other Chronic or Potentially Disabling Conditions Segment of the MBSF contains summarized clinical information for all beneficiaries included in the requested cohort. This file includes two variables for each of the conditions. These variables include: 1) a yearly indicator – which indicates whether each of the condition definitions was met during the respective time period ending December 31, YYYY, and 2) a first occurrence date – which indicates the date the beneficiary was first identified as having met the specifications for the condition (note: 1999 is the earliest year that will appear in this field). Please note that these condition fields are defined by looking at a pattern of medical care utilization, as determined by Medicare FFS claims<sup>4</sup>. See <https://www.ccwdata.org/web/guest/data-dictionaries> for the record layout.

---

<sup>4</sup> A version of these condition variables that examines both Medicare and Medicaid enrollment and claims data is available in MMLEADS. This information may be particularly helpful for projects involving the dually enrolled (Medicare-Medicaid) population.

The two different Condition Segments of the MBSF are constructed each year, based on the specified reference period for each condition. The condition variables for each of the conditions have values which signify whether the pattern of utilization (i.e., FFS claims) indicated the presence of the condition for the beneficiary during the surveillance period ending with the last month of the reference period (e.g., December 2015 for the yearly indicators in the 2015 Conditions Segment; June 2015 for the mid-year indicators in the 2015 Conditions Segment). It is important to note that claims prior to the reference year (e.g., 2015) may have been examined to make this determination, if the CC definition was a 2- or 3- year condition (e.g., diabetes, HF, Alzheimer's disease). Refer to the CC definitions document for more details regarding reference periods and clinical specifications for individual CC definitions (see <https://www.cwdata.org/web/guest/condition-categories>).

## Chapter 3: Calculating Population Statistics

### A. Measures

The CCW data files will likely contain a slightly broader cohort of beneficiaries and claims than you will need for your particular study. Two key considerations must be made before identifying a final study cohort.

#### 1. Numerators

The group of beneficiaries represented in the numerator typically consists of those with particular clinical conditions – or those who have received certain services. Determining who to include in your study group is a crucial first step in study design. The CCW allows for much flexibility in terms of being able to easily refine your selection criteria. There are three<sup>5</sup> types of condition indicator variables to consider:

- **Yearly Indicator** – The first option is to use the yearly indicator for the condition of interest, assuming you have an interest in one or more of the 27 predefined CCW conditions or the Chronic or Potentially Disabling Conditions. These indicators, which consist of a separate field for each condition (e.g., AMI, ALZH, HIP\_FRACTURE), are located in the CCW Chronic Condition Segment of the MBSF and/or the CMS Other Chronic or Potentially Disabling Conditions Segment (note that variable naming convention in the latter file uses “\*\_MEDICARE” in the file name; e.g., AUTISM\_MEDICARE). The value within each field indicates whether the beneficiary received services during the time frame to indicate treatment for the condition (i.e., based on the FFS administrative claims pattern, the beneficiary likely is being treated for the condition – or not). The same variable also indicates whether the beneficiary was able to be observed for the full surveillance period – or until the date of death (i.e., based on Medicare coverage criteria – full Part A and Part B coverage, and no HMO).

Each yearly indicator uses December 31 as the end of the reference year (e.g., 2015 yearly indicator for an algorithm with one-year reference period includes services between 01/01/15 – 12/31/15). The following are valid values for the yearly indicator for each of the conditions:

0 = Beneficiary did not meet claims criteria or have sufficient FFS coverage

1 = Beneficiary met claims criteria but did not have sufficient FFS coverage

---

<sup>5</sup> For the CMS Other Chronic or Potentially Disabling Conditions there are only two variables per condition per year (midyear indicator is not available).

- 2 = Beneficiary did not meet claims criteria but had sufficient FFS coverage
- 3 = Beneficiary met claims criteria and had sufficient FFS coverage

Using information from the four values (0 - 3) in each of the condition indicator fields, researchers may efficiently extract their cohort of interest. Condition indicator values of “3” mean that the pattern of utilization indicates the beneficiary was being treated for the condition, and the beneficiary had Medicare Part A and B coverage – and no HMO coverage for the entire surveillance period – or until death (i.e., anywhere from 1 to 3 years, depending on the condition of interest).

The value of “1” means that the pattern of claims indicates the beneficiary is being treated for the condition – however, the subject was not able to be observed for the full surveillance period. This limited surveillance period could be due to new accretion into the Medicare program (i.e., beneficiaries who became newly eligible), a break in Part A or B coverage, or one or more months of managed care coverage. CCW routinely delivers to researchers claims and/or assessment information for values of three (3) and one (1) for the requested condition, unless the researcher specifies otherwise.

The other two potential values in this field may indicate *absence* of the condition during the reference period (condition indicator for the condition = 2 or 0). The twos [2’s] had coverage throughout the full surveillance period, the zeroes [0’s] did not. In both cases there were no claims to indicate current treatment for that particular condition.

Researchers will need to determine whether to include only beneficiaries with a condition and the full surveillance period (i.e., the threes [3’s]) – or whether it is valuable to retain some or all of the beneficiaries who appear to have the condition – even though coverage for the full surveillance period may be lacking (i.e., the ones [1’s]). Through merging the condition Summary File with the Beneficiary Summary File (using BENE\_ID) one can efficiently determine level of coverage during the year for each of the beneficiaries. More information regarding how to make cohort selections based on Medicare coverage criteria is found below, in the **Denominators** section of this paper.

- **Mid-Year Indicator**<sup>6</sup> - If researchers are interested in a surveillance period that does not correspond to a calendar year, the mid-year indicators may be of interest. Like the yearly indicators, the mid-year indicators are located in a separate field for each condition (e.g., AMI\_MID, ALZH\_MID, HIP\_FRACTURE\_MID). The definitions for each of the conditions are the same as for the yearly indicators – including the reference period. For the mid-

---

<sup>6</sup> The midyear indicator is not available for the CMS Other Chronic or Potentially Disabling Conditions.

year indicators, the reference periods end on June 30 of the year – rather than on December 31 of the year.

- **First Occurrence Date** – Another option for determining who has a condition is to ask whether the beneficiary *ever* had the condition. These “ever” fields, a separate one for each of the 27 CCW conditions, can also be found in the Chronic Condition Segment of the MBSF; the “ever” fields for the CMS Other Conditions can be found in the Other Chronic or Potentially Disabling Conditions Segment of the MBSF. The presence of a date in the field (e.g., AMI\_EVER; formatted as YYYYMMDD) – indicates the date the beneficiary first met the *clinical* criteria of the algorithm (no coverage criteria applied), with the earliest possible date of 19990101. A null value indicates this diagnosis has never been met since enrollment in Medicare.

## 2. Denominators

Determining who is “at risk” for the events of interest is an important next step in the research process. The opportunity to observe FFS claims which may indicate the presence (or confirm the absence) of treatment for the condition can only occur if there is some period of FFS coverage exposure. Beneficiaries have a variety of Medicare coverage options, and may not have FFS for the full surveillance period of interest. It is up to the researcher to decide how much coverage is sufficient to be included in the study.

*We wish to caution researchers against using the total number of beneficiaries included in the MBSF (or even a particular subsample such as the 5% random sample) as a denominator for any condition prevalence calculation. This number is an overestimate of the number of beneficiaries with FFS and at risk for the condition at any point in time (refer to **Results** section of this paper for an illustration). Please be deliberate in making your denominator selection in order to produce accurate rates.*

There are several variables in the MBSF (A/B/C/D Segment) that can be used to specify which beneficiaries to include in the study. For example, cohorts may be limited to certain ages, geographic locations, or a particular gender. Researchers may also want to retain beneficiaries who meet certain coverage criteria. Some of these variables include:

- Coverage variables - whether the beneficiary was covered by Part A and/or Part B during a particular month
- State buy-in variables - a proxy for dual eligibility in Medicaid and Medicare (monthly variables BENE\_MDCR\_ENTLMT\_BUYIN\_IND\_01 - 12)



- Managed care variables - whether the beneficiary had managed care coverage (monthly variables BENE\_HMO\_IND\_01 - 12)
- Date of date (BENE\_DEATH\_DT).

The MBSF and all Segments are created on an annual basis, therefore, should be requested for each calendar year of interest. Each MBSF A/B/C/D Segment provides the demographic and coverage information for a given year. For researchers interested in requesting multiple years of data, multiple years of the MBSF should also be requested (e.g., if three years of claims data are requested, then three years of MBSF should be requested).

## **B. Analysis of Rates**

There are many different types of rates that can be constructed using the condition indicator variables. This paper is primarily intended to discuss options for calculating population-based rates for chronic conditions.

### **1. Description of Options**

The types of rates described should not be interpreted as incidence rates. They are the prevalence of FFS treatment for a clinical condition at a point-in time, using treatment/receipt of services as a proxy for having the condition of interest (e.g., claims-based point-prevalence).

These analytic options all include some examination of the extent and duration of Medicare coverage. In general, researchers may wish to look at enrollment in Part A and Part B, since most of the condition definitions include either or both types of services (e.g., a combination of inpatient, outpatient and/or Carrier claims). However, for a couple of conditions (i.e., AMI or hip fracture) one inpatient claim is sufficient to indicate presence of the condition. For these conditions, the researcher should ponder whether to include subjects without Part B coverage (e.g., for studies related to follow-up care, researchers may wish to retain only subjects who are also enrolled in Part B, or include only subjects with Part A coverage for a study related to inpatient care).

In addition, since FFS claims are central to the algorithms used by CCW to identify beneficiaries with conditions, the researcher may wish to allow little or no managed care coverage in the surveillance period. This will allow for sufficient opportunity to see a Medicare claim, indicating the utilization of interest.

Researchers interested in calculating population-based rates (as opposed to rates of events within the study cohort – such as readmission or mortality rates), may find it desirable to use

the random 5% sample (or some other randomly-selected population), as the file sizes are more manageable and require the minimum data necessary to perform the research study.

The beneficiaries represented in the rate calculations depend on how restrictive or lenient the researcher wishes to be in terms of inclusion criteria. Four denominator options are contrasted below. The same coverage restrictions should be applied to both numerator and denominator (it is assumed the researcher selects the population first – then, from this sample, determines who has the condition[s] of interest):

- **Full FFS coverage** – The beneficiary has Part A, Part B and no HMO coverage for the full surveillance period (or until the time of death, if the beneficiary died in the year). This type of restriction limits analysis to those with full FFS coverage, which means that some beneficiaries known to have the condition of interest (i.e., because there are sufficient claims to indicate presence of treatment for the condition) may be excluded from the study. This is the most restrictive type of coverage option. This type of rate is the simplest to compute in CCW, as data values already exist to provide numerators and denominators for the condition variables. However, the researcher should not presume that this cohort is either representative of the Medicare population as a whole, or of the Medicare FFS population, as beneficiaries with full coverage may not be “typical” or representative of all Medicare consumers.
- **Partial FFS coverage** – The beneficiary has some Medicare Part A and/or B coverage (or much coverage, depending on how the coverage criteria are specified), and may or may not have some managed care coverage. This option allows for a break in coverage, and is less restrictive than the “full FFS coverage” option above. A common recommendation is to allow for a one month break in coverage per year of surveillance; in fact, by requiring 11 or 12 months of coverage (i.e., full or nearly full FFS), you will exactly match the rate calculations on the CCW website. This is an attractive option to avoid losing any/many cases with the condition of interest (i.e., known cases, as indicated in claims) due to the occurrence of partial FFS coverage.
- **Point in time FFS coverage** – The beneficiary has FFS coverage during the month of interest – (e.g., for July, the midpoint of the year). This is an appealing option for identifying enrollment and disease burden for a typical point in time (i.e., a month).
- **Person years with FFS coverage** – This option allows for each beneficiary with any FFS coverage to count toward the denominator, for however many months (or proportion of a year) they have the coverage of interest. Using a denominator such as this, all cases are counted in the numerator, and their corresponding time “at risk” is included in the

denominator. This option is attractive since beneficiaries are “under FFS observation” for various lengths of time. Technically, rates are not produced using this type of calculation – rather the extent of illness is expressed at a ratio of cases to the time at risk.

## 2. Calculation of Options

Several analytic steps are required to accurately capture prevalence rates using the CCW data. Analytic guidance for calculating these four different types of prevalence rates are described. The results of each type of analysis are presented in the **Results** section of this paper so researchers may assess the empirical difference in numerators, denominators, and rates that these analytic variations produce.

We share code examples below to illustrate various calculation methods. Throughout, we limit our analyses to the CMS random 5% sample; our examples use 2015 data files. Since the CCW data files are very large, we recommend retaining as few rows of data and as few variables as possible to allow the programs to run efficiently. You may wish to review additional CCW Technical Guidance papers to learn more about options for specifying analytic cohorts; see, for example “Getting Started with CMS Medicare Administrative Research Files – A Technical Guidance Paper” (<https://www.cwdata.org/web/guest/technical-guidance-documentation>).

Using the MBSF for the year of interest, retain only beneficiaries selected as part of the random 5% sample for that year (i.e., using the SAMPLE\_GROUP variable – where the value = ‘01’ or ‘04’). Keep only the subset of variable of interest. Next, merge this restricted MBSF A/B Segment with the CCW Conditions Segment and/or Other Chronic or Potentially Disabling Conditions Segment (linkage variable = BENE\_ID) for the same year and retain only the beneficiaries in the population selected from the MBSF. Now, you are ready to examine prevalence rates using this combined file, which we assume is named “**Prevalence20XX**” in the sample SAS<sup>®</sup> code provided later in the paper.

- **Full coverage** – This is the simplest type of prevalence rate to calculate using the CCW data. For this option, you may simply use variables from the MBSF Conditions Segment for the year of interest (note: although there is not a need to link with the MBSF for the purpose of incorporating coverage information, we recommended that the cohort be limited to the random 5% sample first – as described above).

Once you have identified your CC variable of interest, this rate is calculated by using certain values within the data element.

$$\text{Denominator} = [2's] + [3's]$$
$$\text{Numerator} = [3's]$$

For example, if we are interested in AMI as a CC, once we have subsetting the data to obtain a strict 5% file, we simply need to look at the AMI variable (which is called AMI) – and keep the threes [3's] and twos [2's] (both indicate full Medicare FFS coverage – and the presence [3's] or absence [2's] of the condition). The prevalence rate = ratio of (3) / (3 + 2). Note that this calculation is identical to using the code in example for Partial coverage (see Code example 1, below) if you edit the code to require a full 12 months of coverage.

- **Partial coverage** – For this option, our example allows for a one month break in Part A and/or Part B coverage, and up to one month of managed care coverage per year of surveillance (e.g., beneficiaries with 11 or 12 months of Part A, B, no HMO for one year conditions are retained; similarly, for 2-year conditions you might retain beneficiaries with 22 out of 24 months of FFS coverage). See Code example 1, below.

Another important step for this calculation is to keep beneficiaries who were fully covered until the time of death (or covered for all but a month prior to the time of death). Failure to have an extra analytic step to include those who died will likely result in an undercount for both the numerator and denominator in meaningful ways (e.g., you would fail to count fatal AMIs which occurred in any month other than at the end of year; a death could result in very few months of Medicare coverage for the beneficiary).

Using the **Prevalence20XX** file, you will need to combine information from three key variables.

- The information regarding Part A and Part B coverage for each month during the year is contained in a series of 12 variables (one for each month) called MDCR\_ENTLMT\_BUYIN\_IND\_01 – 12. For this example, we count only the beneficiaries with Part A and B coverage (regardless of whether they have the state buy-in). For this series of variables, we want to keep values of 3 (indicating the beneficiary has Part A and B) and C (indicating the beneficiary has state buy-in for Part A and B).
- Next, we assume you will also want to *exclude* beneficiaries who have more than one month of managed care coverage during the time frame. Managed care information is contained in a series of 12 variables (one for each month) called HMO\_IND\_01-12. Any value other than zero (0) indicates that there was some type of managed care coverage during the month. Please note that a value of four (4) in

the HMO indicator variables is used to identify beneficiaries included in a FFS demonstration project (in 2012 and forward) - and FFS claims are available in the data files for these beneficiaries. Hence, we do not exclude these beneficiaries (the fours [4's]) from consideration in our algorithms.

- The information regarding which beneficiaries died during the year, and when the death occurred is located in the BENE\_DEATH\_DT field.

To construct the denominator and numerator which include beneficiaries with partial coverage, and those who were covered until the time of death, we include an example of analytic code using SAS<sup>®</sup> programming language. You may adapt this code to use whatever software you prefer.

**Code Example 1.** Define Sample of Beneficiaries with Full or Partial Medicare Fee-For-Service (FFS) Coverage Using Variables from MBSF

Define a sample of Medicare FFS beneficiaries with a specified duration of A and B FFS coverage, or coverage until the month of death. The first portion of the code uses the monthly coverage variables to require at least 11 months of FFS coverage; the second portion of code identifies beneficiaries with the condition of interest (e.g., AMI). The input data source is the Master Beneficiary Summary File – only key variables from the A/B/C/D segment and the CCW Conditions Segment. All of our examples use data from 2015 (note: the variables and variable names are the same over time).

The following SAS<sup>®</sup> code assumes that you previously created a data file to use as our input data (called prevalence20xx); it should have merged Medicare coverage and Condition variables for your population of interest, keeping only the necessary study variables.

We begin by developing a cohort to use for population statistics (called *temp*) by using Medicare coverage specifications. Then you may use these variables to calculate rates for your population (e.g., FFS only).

```
data temp;
set ccw.prevalence20xx;
*note this input file is a merged MBSF A/B/C/D and Conditions
File Segment;
* determine # months of Part A, B and no HMO coverage;

array MemberMos_AB (12)
mdcr_entlmt_buyin_ind_01 - mdcr_entlmt_buyin_ind_12;
array MemberMos_noHMO (12) hmo_ind_01 - hmo_ind_12;
array Member_FFSMos (12) Member_FFSMos01 - Member_FFSMos12;
```

```
do i= 1 to 12;
  if MemberMos_AB(i) in ('3','C') and MemberMos_noHMO(i) in
  ('0','4')then Member_FFSMos(i)=1;
  else if MemberMos_AB(i) NOT in ('3','C') or
  MemberMos_noHMO(i) NOT in ('0','4')then Member_FFSMos(i)=0;
  Member_Mos=sum(of Member_FFSMos:);
end;

* note - determine who had 11 or 12 months coverage or coverage
until month prior to death;

if (bene_death_dt=. and Member_Mos in (11,12)) or
(bene_death_dt~=.
and month(bene_death_dt)<=Member_Mos+1 and Member_mos~=0) then
Partl_Cov=1;
else Partl_Cov=0;

* note - bring in numerator information for AMI - keep both 3s
and 1s;

if Partl_Cov=0 then AMI_PartRT=.;
else if Partl_Cov=1 and (ami= 0 or ami=2)then AMI_PartRT=0;
else if Partl_Cov=1 and (ami= 1 or ami=3)then AMI_PartRT=1;

label
Partl_Cov = '11 or 12 months FFS no HMO - except for those who
died'
Member_Mos = 'Total Member months of A B and No HMO - per bene'
AMI_PartRT = 'Had AMI - partial coverage';
run;
```

To compute the partial coverage rate we simply aggregate the AMI\_PartRT variable created in the data step using the means procedure. The code below will produce three outputs: N will be the rate denominator, SUM will the rate numerator, and MEAN will be the rate.

```
proc means data=temp N SUM MEAN;
var AMI_PartRT;
run;
```

With a few minor changes, this analytic code can be used as the basis for code which can specify the other types of cohorts discussed in this paper. You can easily modify this SAS® code to fit your own denominator specifications (e.g., - two month break in coverage; not requiring Pt B coverage – only A coverage).

*Note:* The condition statistics on the CCW website (i.e., Medicare Table B.2) use the Partial coverage algorithm and require 11 or 12 months of coverage for inclusion in the denominator (<https://www.ccwdata.org/web/guest/medicare-tables-reports>).

- **Point in time coverage** – For this type of a denominator, we need to determine which Medicare beneficiaries were alive and had Part A and Part B coverage, and no HMO coverage, during our month of interest. For our example, we use the midpoint of the year, and assess who has coverage in July of our reference year (2015).

If this is the only denominator you have an interest in, you may simply construct your denominator using two key variables. The first is MDCR\_ENTLMT\_BUYIN\_IND\_07 (the 07 extension on this variable corresponds with the 7th month of the year, July). As in the example above, we will count only the beneficiaries with Part A and B coverage (i.e., values of 3 and C for this variable). Next, we want to exclude beneficiaries who have managed care coverage during the same month. The variable to use is HMO\_IND\_07 (any value other than 0 = managed care coverage; we also want to retain the 4's since FFS claims are available).

Once you have a subset of beneficiaries who meet the denominator criteria, again you would count your cases (i.e., your numerator) as the threes [3's] and ones [1's] for your condition of interest. All others in your denominator do not have evidence (FFS claims) indicating treatment for the condition. Adding the following lines of code to the data step above will create the needed variable.

**Code Example 2.** Define Sample of Beneficiaries with Medicare Fee-For-Service (FFS) Coverage at a Point in Time

The following SAS<sup>®</sup> code assumes that you are working with source data that includes the MBSF A/B/C/D and CCW Conditions Segments. We determine which beneficiaries were enrolled in FFS for the particular month (FFS for July only), then we identify those with the condition of interest (which is an annual variable).

```
if mdcr_entlmt_buyin_ind_07 in ('3','C') and hmo_ind_07 in  
( '0','4') then Member_FFSMos07=1;  
else if mdcr_entlmt_buyin_ind_07 NOT in ('3','C') or hmo_ind_07  
NOT in ('0','4') then Member_FFSMos07=0;
```

```
if Member_FFSMos07=0 then AMI_PtTimeRT=.;  
else if Member_FFSMos07=1 and (ami= 0 or ami=2)then AMI_PtTimeRT  
=0;  
else if Member_FFSMos07=1 and (ami= 1 or ami=3)then AMI_PtTimeRT  
=1;
```

Again to compute the point in time coverage rate we simply aggregate the AMI\_PtTimeRT variable created in the data step using the means procedure.

```
proc means data=temp N SUM MEAN;  
var AMI_PtTimeRT;  
run;
```

- **Person years with coverage** – For this type of denominator, the objective is to ascertain Medicare FFS member years of coverage (i.e., cumulative member months at risk, divided by 12). For a one-year condition (AMI) we will accumulate 12 months of coverage information for each beneficiary in the 5% sample (note: for a 2 year condition you would want to accumulate 24 months of coverage, etc.). Then, for the purposes of comparing methods for calculating prevalence, we divide by 12 to obtain an “average” member months at risk, and use this for our denominator. This is similar to the denominator calculation for a traditional “period prevalence” type of rate.

You will need to create “counter” variables which accumulate the number of months each beneficiary meets your coverage criteria. Using our same definition of coverage (Part A and Part B with no HMO) – we count the number of months this definition is met. We illustrate this process using Code example 3, below.

For the numerator, we need to determine how often events occur. Using this denominator, we count the threes [3’s] and ones [1’s] for the condition for the identified time period (note: for this method there is not the need to link beneficiaries – but to count months at risk, and count events). Please use caution for interpreting this type of a rate. It is not technically incidence, but rather indicates the number of beneficiaries at risk who had a treatment event during the time period of interest.

**Code Example 3.** Determine Total Member Months of Medicare Fee-For-Service (FFS) Coverage, and Corresponding Condition Numerator

This time add the following lines of code to the data step.

```
Member_Years = Member_Mos/12;  
if (ami= 0 or ami=2)then AMI_Flag=0;
```



```
else if (ami= 1 or ami=3)then AMI_Flag=1;
```

To compute the person years coverage rate we compute the sum the Member\_Years (denominator) and AMI\_Flag (numerator) variables created in the data step using the means procedure.

```
proc means data=temp SUM;  
var Member_Years AMI_Flag;  
run;
```

## Chapter 4: Results

The four methods for calculating prevalence, described above, produce somewhat different rates. The fourth (4<sup>th</sup>) method described is formatted in Table 3 as a reminder that this type of a statistic is technically different from the others. In the fourth method, a particular cohort is not identified; rather the number of events for person time at risk is displayed.

**Table 3.** Four Methods for Calculating Prevalence of AMI in 2015 (5% random sample)

	Denominator	Numerator	Percent
No coverage restrictions*	2,918,854	15,768	0.54%
1. Full FFS coverage	1,594,530	14,201	0.89%
2. Partial FFS coverage	1,611,855	14,321	0.89%
3. Point in time FFS coverage	1,653,733	13,125	0.79%
4. <i>Person years with FFS coverage</i>	<i>1,650,997.8</i>	<i>15,768</i>	<i>0.96%</i>

*\*This is not a viable method (it is displayed for comparative purposes only)*

It is apparent that the first row Table 3 (shaded, in order to contrast this from the more deliberate and accurate rate computations) has a denominator much larger than the other viable methods – and a correspondingly low rate. The denominator is the sample size for the entire random 5% sample for 2015, and the numerator includes all the threes (3’s) and ones (1’s) for the condition. ***We do not recommend ever attempting to describe prevalence of a condition using this technique***, as beneficiaries with no FFS coverage are included (i.e., those who are not “at risk” for having a FFS claim indicating presence of the condition).

Determining which rate is the correct rate to use depends on the research question and purpose of the analysis. For example, you may wish to include only those with full FFS coverage in your study if it is essential to obtain a thorough description of all services used by beneficiaries with a certain condition. This maximizes the FFS surveillance opportunity, however, you lose some of the known cases in the cohort (i.e., none of the ones [1’s] are included).

Allowing a partial break in coverage allows you to keep more known cases. If we understand that full FFS coverage may not necessarily be “typical” of all Medicare beneficiaries, then this type of rate may be more generalizable. The patterns of care, costs and outcomes for this cohort should be fairly completely ascertained, as the surveillance period is still extensive.

Point in time coverage is a method often appealing to those trying to extrapolate rates to the entire population – FFS beneficiaries for a typical month (and costs for these beneficiaries for a typical month, etc.).

Being able to include all known events is part of the appeal of the person-time method for calculating these rates – and every month of FFS “at risk” is counted. This method is useful for the purposes of constructing prevalence-like ratios; however, it is not particularly helpful if the researcher wishes to examine claims or patterns of care for those in the denominator. A single month of coverage results in beneficiary information being counted, yet there is not a strictly defined cohort.

### **Next Steps**

Perhaps the purpose of your study goes well beyond simply determining the prevalence of the condition of interest. Once you have defined your denominator, you may proceed to the rest of your analysis (e.g., assessing utilization, health outcomes, etc.) – including only this subset of beneficiaries (i.e., your cohort of interest).

## Chapter 5: Generalizing These Methods

Some conditions in the CCW require a two or three year surveillance period – (e.g., diabetes requires two years and Alzheimer’s requires three years). The methods described herein can be generalized to these conditions. For prevalence calculation method #1 (Full FFS coverage), you would simply need to look at the yearly indicator variable for the condition of interest. By definition, the 3’s and 2’s indicate that the beneficiary had full coverage for the surveillance period (not just the calendar year represented by the MBSF). For the other methods, you would need to obtain multiple years of the MBSF A/B/C/D Segment in order to accumulate months of coverage for the entire duration of the surveillance period. During the data request process, you are able to specify which years of the MBSF are desired.

Our examples used the yearly condition indicator variable for all numerator calculations. You may choose to use the mid-year variable (e.g., AMI\_MID) or the first occurrence date (e.g., AMI\_EVER) for numerator determination. Some conditions included in the CCW may not require active treatment, yet it is helpful to know the disease history (e.g., breast cancer, stroke) so that a “clean” control group can be obtained. The “ever” variables (i.e., first occurrence date) are designed for this purpose.

An additional potential use of the “ever” variables is to look at the population subset with a recurrence of disease. For example, perhaps you are interested in studying stroke or breast cancer. Beneficiaries identified as having the condition using the current year variable (called STROKE\_TIA or CANCER\_BREAST) may consist of subgroups who have never had the condition before – and those who have a prior history of the condition (i.e., perhaps the current year actually indicates a recurrence of the condition rather than an incident case). These subgroups can be identified by determining whether the ever date was prior to the year of the data file (e.g., if the date value in the STROKE\_TIA\_EVER variable is prior to 2015, then the stroke indicated in the 2015 data file is a recurrent stroke). The earliest possible ever date for anyone in the CCW is January 1, 1999, although some beneficiaries became eligible for Medicare before then. For beneficiaries who joined Medicare after that date, their ever dates will not precede the start of their Medicare coverage (i.e., the COVSTART variable in the MBSF A/B/C/D Segment). *Investigators are cautioned that the absence of an ever date that precedes the current year does not mean the person was never treated for the disease; it simply means that we have no Medicare claims history (from 1999 forward) indicating the presence of the condition.*

For conditions that only require an inpatient diagnosis code in order to be classified according to the condition definition, you may wish to construct more lenient coverage criteria than we discussed herein. For example, for hip fracture (variable called HIP\_FRACTURE), the presence of

one inpatient diagnosis code is sufficient for meeting the claims-based definition. You should determine whether you wish to look at cohorts with Part A coverage, and/or those with Part B coverage. By keeping the coverage criteria as lenient as possible, more of the known cases are retained. The SAS® code displayed above can be easily adapted to accommodate these scenarios.

## **A. Limitations**

The condition definitions in CCW are claims-based definitions – determined by documentation of receipt of treatment for the condition of interest in the Medicare FFS claims data. As a result, population prevalence rates derived from this data source may differ somewhat from prevalence rates constructed from other data sources – particularly those derived from survey data or other types of clinical data.

The claims used to make the condition determinations are for Medicare FFS only. As a result, there is missing data in the CCW due to managed care coverage. Since claims for most services provided to Medicare beneficiaries in managed care do not reach the claims data files, the CCW Medicare claims should be viewed as providing utilization information primarily for the FFS population. The managed care population may differ in important ways from the FFS population (e.g., they could potentially be younger and healthier). Population-level generalizations made using CCW data should be made with caution.

Researchers have a variety of hypotheses and objectives. The intent of this paper is not to be prescriptive, but rather descriptive of some useful tools for refining cohorts and calculating rates. The objective is to make it easy for you to accomplish your study objectives – and to ensure you are able to do so with a thorough understanding of the data available from the CCW. Using the methods described in this paper, we can gain a better understanding of the magnitude of chronic conditions, and the effect on the population, through appropriate and accurate data analysis techniques.

## Chapter 6: Further Assistance with Medicare Administrative Claims Data

The Research Data Assistance Center (ResDAC) offers free assistance to those using Medicare data for research. The ResDAC website provides links to descriptions of the CMS data available, request procedures, supporting documentation, workshops on how to use Medicare data and other helpful resources. Visit the ResDAC web site at (<http://www.resdac.org>) for additional information.

ResDAC is a CMS contractor and requests for assistance in the application, obtaining, or using the CCW data should first be submitted to ResDAC. Investigators can reach ResDAC by phone at 1-888-973-7322, e-mail at [resdac@umn.edu](mailto:resdac@umn.edu), or online at (<http://www.resdac.org>).

In the event that a ResDAC technical advisor is not able to answer the question, the technical advisor will direct the investigator to the appropriate person. If additional CMS data (data not available from the CCW) is required to meet research objectives, or the investigator has any questions about other data sources, the investigator can review all available CMS data by visiting the ResDAC website and contact ResDAC for further assistance.

[www.ccwdata.org](http://www.ccwdata.org)  
Email: [CMSdata@gdit.com](mailto:CMSdata@gdit.com)  
Phone: 1-866-766-1915

## Appendix A: List of Acronyms and Abbreviations

<b>Acronym</b>	<b>Definition</b>
AMI	Acute Myocardial Infarction
BASF	Beneficiary Annual Summary File
BSF	Beneficiary Summary File
BETOS	Berenson-Eggers Type of Service
CC	Chronic Condition
CCW	Chronic Conditions Data Warehouse
CMS	Centers for Medicare & Medicaid Services
COPD	Chronic Obstructive Pulmonary Disease
FFS	Fee-for-Service
HIV/AIDS	Human Immunodeficiency Virus / Acquired Immunodeficiency Disorder
HMO	Health Maintenance Organization
MAX	Medicaid Analytic eXtract
MBSF	Master Beneficiary Summary File
MMA	Medicare Modernization Act of 2003
MMLEADS	Medicare-Medicaid Linked Enrollee Analytic Data Source
PTSD	Post-Traumatic Stress Disorder
ResDAC	Research Data Assistance Center