

Brian O'Donnell, PhD

Kathy Schneider, PhD

Debbie Dean, MS

---

**Buccaneer Computer  
Systems & Service, Inc.**  
Software Development &  
Informatics Division

1401 50<sup>th</sup> Street, Suite 200  
West Des Moines, IA 50266

(515) 645-3010

[www.ccwdata.org](http://www.ccwdata.org)

# **CMS Chronic Condition Data Warehouse**

## **Technical Guidance for Researchers Calculating Population Statistics**

*A technical guidance paper developed under  
contract with the Centers for Medicare &  
Medicare Services*

**September 2008**

## Introduction

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 for a random 5% sample of Medicare beneficiaries – and for 100% of the Medicare FFS population from 2005 forward. The CCW data are linked by a unique, unidentifiable beneficiary key, which allows researchers to analyze information across the continuum of care.

The CCW data files were designed to support a variety of research objectives. The 21 predefined CCW 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 chronic condition determinations (i.e., an indicator that the beneficiary received a service or treatment for the condition of interest).

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. These 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 on the CCW website 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 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.

## Objectives

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.

## Contents of CCW

The sample of beneficiaries in CCW includes data for the strict 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 100% of Medicare-enrolled beneficiaries, depending on the years of interest. Data for all eligible beneficiaries are contained within the CCW database (i.e., 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 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 the 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. Two different versions of read-in statements are routinely provided – for both the Beneficiary Summary File and the Chronic Condition Summary File. The SASv6 file (e.g., `beneficiary_summary_file_read_v6.sas`) contains traditional short SAS<sup>®</sup> names, which may be appealing for researchers who have worked with historical CMS data. The SASv8 read-in statements (e.g., `beneficiary_summary_file_read_v8.sas`) take advantage of newer features in SAS<sup>®</sup>, which allow 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.

## Methods

### A. Data Files

Two key CCW data files allow researchers to easily determine which subjects to keep in their study. These useful files are distributed routinely with approved data requests – at no extra charge to the requestor.

- **Beneficiary Summary File** – The Beneficiary Summary File is created annually and contains demographic eligibility and enrollment data for all beneficiaries who are alive and eligible for any part of the year. This file contains a flag that indicates whether a beneficiary was included in the CMS 5% sample for the year (i.e., the cross-sectional annual sample) 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). 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 (see [http://www.ccwdata.org/variables/var\\_beneficiary\\_summary.php](http://www.ccwdata.org/variables/var_beneficiary_summary.php) for the record layout) 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.
- **Chronic Condition Summary File** – The Chronic Condition Summary File contains summarized clinical information for all beneficiaries included in the requested cohort. This file includes a set of three variables for each of the 21 chronic conditions. These three

variables include: 1) a yearly indicator – which indicates whether each of the 21 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 [http://www.ccwdata.org/variables/var\\_chronic\\_condition\\_summary.php](http://www.ccwdata.org/variables/var_chronic_condition_summary.php) for the record layout.

The Chronic Condition Summary File is constructed each year, based on the specified reference period for each condition. The three chronic condition (CC) variables for each of the 21 CCs 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 2005 for the yearly indicators in the 2005 CC Summary File; June 2005 for the mid-year indicators in the 2005 CC Summary File). It is important to note that claims prior to the reference year (e.g., 2005) may have been examined to make this determination, if the CC definition was a 2- or 3- year condition (e.g., diabetes, CHF, Alzheimer's). Refer to the CC definitions document for more details regarding reference periods and clinical specifications for individual CC definitions (see <http://www.ccwdata.org/downloads/Chronic%20Condition%20Data%20Warehouse%20Condition%20Categories.pdf>).

## *B. 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.

- **Numerators** The group of beneficiaries represented in the numerator typically consist 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 types of CC indicator variables to consider:

### *1. CC Yearly Indicator*

The first option is to use the CC yearly indicator for the condition of interest, assuming you have an interest in one or more of the 21 predefined CC's. These indicators, which consist of a separate field for each condition (e.g., AMI, ALZH, HIPFRACTURE), are located in the CC Summary File. 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., 2005 yearly indicator for an algorithm with one-year reference period includes services between 01/01/05-12/31/05). The following are valid values for the yearly indicator for each of the 21 CC's:

0 = Neither claims nor coverage met

- 1 = Claims met, coverage not met
- 2 = Claims not met, coverage met
- 3 = Claims and coverage met

Using information from the four values (0 - 3) in each of the CC indicator fields, researchers may efficiently extract their cohort of interest. CC 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 CC, unless the researcher specifies otherwise.

The other two potential values in this field may indicate *absence* of the condition during the reference period (CC 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 CC.

Researchers will need to determine whether to include only beneficiaries with a CC 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 CC 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.

## 2. *Mid-Year Indicator*

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, HIPFRACTURE\_MID). The definitions for each of the CCs are the same as for the yearly indicators – including the reference period. For the mid-year indicators, the reference periods end on June 30 of the year – rather than on December 31 of the year.

## 3. *First Occurrence Date*

Another option for determining who has a CC of interest is to ask whether the beneficiary *ever* had the CC. These “ever” fields, a separate one for each of the 21 CC’s, can also be found in the CC Summary File. 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.

- **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 5% sample as a denominator for any CC 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 Beneficiary Summary File that can be used to narrow which beneficiaries to include in the study. For example, study groups 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\_IND01 - 12)
- Managed care variables - whether the beneficiary had managed care coverage (monthly variables BENE\_HMO\_IND\_01 - 12)
- Date of death (BENE\_DEATH\_DT).

Beneficiary Summary Files are created on an annual basis, therefore, should be requested for each calendar year of interest. Each Beneficiary Summary File provides the demographic and coverage information for a given year. For researchers interested in requesting multiple years of data, multiple years of the Beneficiary Summary File should also be requested (e.g., if three years of data are requested, then three Beneficiary Summary Files should be requested).

### C. Analysis of Rates

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

- **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 CC 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 CCW identification of 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. Since the CCW is an *ever* enrolled 5% data file, in order to extrapolate results for national estimates, the researcher will need to use the CCW Beneficiary Summary File and narrow their analyses to the strict 5% sample.

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):

1. *Full 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 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 21 CCs. 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.
2. *Partial coverage.* The beneficiary has *some* Medicare Part A and/or B coverage (or much, 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 coverage” option above. A common recommendation is to allow for a one month break in coverage per year of surveillance. 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.
3. *Point in time coverage.* The beneficiary has 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).
4. *Person years with coverage.* This option allows for each beneficiary with any 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.

- **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.

First, retain the strict 5% sample. Using the Beneficiary Summary File for the year of interest, retain only beneficiaries selected as part of the random 5% sample for that year (i.e., using the FIVE\_PERCENT\_FLAG variable – where the value = Y). Next, merge this restricted Beneficiary Summary File with the CC Summary File for the year of interest by BENE\_ID and retain only the beneficiaries in the restricted Beneficiary Summary File. Now, you are ready to examine prevalence rates using this combined file, named “Prevalence20XX” in the sample SAS® code provided later in the paper.

1. *Full coverage.* This is the simplest type of prevalence rate to calculate using the CCW data. For this option, researchers may simply use variables from the CC Summary File for the year of interest (note: although there is not a need to link with the Beneficiary Summary File for the purpose of incorporating coverage information, it is 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.

Denominator = [2's] + [3's]  
Numerator = [3's]

For example, if I am interested in AMI as my CC, once I have subsetted the data to obtain a strict 5% file, I 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 coverage – and the presence [3's] or absence [2's] of the condition). The prevalence rate = ratio of (3) / (3 + 2).

2. *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).

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 bene\_medcr\_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 bene\_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 2005 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® programming language. Researchers may adapt this code to use whatever software they prefer.

```

data temp;
set ccw.prevalence20xx;
*note this is a merged CC Summary and Bene Summary file;
* note - determine # months of Part A, B and no HMO coverage;
array MemberMos_AB (12)
bene_mdcr_entlmt_buyin_ind_01 - bene_mdcr_entlmt_buyin_ind_12;
array MemberMos_noHMO (12) bene_hmo_ind_01 - bene_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 coverage until one 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. Researchers can easily modify this SAS® code to fit their own denominator specifications (e.g., - two month break in coverage; not requiring Pt B coverage – only A coverage).

3. *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 (2005).

If this is the only denominator you have an interest in, you may simply construct your denominator using two key variables. The first is `bene_mdcr_entlmt_buyin_ind_07` (the 07 extension on this variable corresponds with the 7<sup>th</sup> 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 `bene_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.

```
if bene_mdcr_entlmt_buyin_ind_07 in ('3','C') and bene_hmo_ind_07 in
('0','4')then Member_FFSMos07=1;
else if bene_mdcr_entlmt_buyin_ind_07 NOT in ('3','C') or
bene_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;
```

4. *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.

For the numerator, one needs to determine how often events occur. Using this denominator, researchers would count the threes [3’s] and ones [1’s] for the CC 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.

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;
```

## Results

The four methods for calculating prevalence, described above, produce somewhat different rates. The fourth (4<sup>th</sup>) method described is formatted in *italics* in Table 1 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 1. Four Methods for Calculating Prevalence of AMI in 2005**

	Denominator	Numerator	Percent
No coverage restrictions*	2,232,528	17,861	0.80
1. Full coverage	1,633,165	16,594	1.02
2. Partial coverage	1,648,175	16,681	1.01
3. Point in time coverage	1,678,332	14,898	0.89
<i>4. Person years with coverage</i>	<i>1,676,003</i>	<i>17,861</i>	<i>1.06</i>

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

It is apparent that the first row in Table 1 (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 2005, 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, it may be desirable to include only those with full FFS coverage in the 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, the researcher loses 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 the researcher 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).

## **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 coverage), researchers would simply need to look at the CCW 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 CC Summary File). For the other methods, multiple years of the Beneficiary Summary File would need to be obtained in order to accumulate months of coverage for the entire duration of the surveillance period. During the data request process, researchers are able to specify which years of the Beneficiary Summary File are desired.

Our examples used the CC yearly indicator variable for all numerator calculations. Researchers 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 for researchers 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.

For conditions that only require an inpatient diagnosis code in order to be classified according to the CCW definition, researchers may wish to construct more lenient coverage criteria than we discussed herein. For example, for AMI, the presence of one inpatient diagnosis code is sufficient for meeting the claims-based definition. Researchers should determine whether they 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<sup>®</sup> code displayed above can be easily adapted to accommodate these scenarios.

## Limitations

The CC definitions in CCW are claims-based definitions – determined by documentation of receipt of treatment for the condition of interest. 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 CC 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 researchers to accomplish their study objectives – and to ensure they 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.