Instructions for the SAS READIN Files for the *DE-SynPUF*

This section provides a summary of how users can access the Centers for Medicare and Medicaid Services (CMS) *Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use Files* (*DE-SynPUF*). These files include fully synthetic data that are designed to be a realistic representation of what the actual data are like, but all the data in these files are for fictional Medicare beneficiaries and their fictional Medicare claims.

DE-SynPUF Subsample Files:

- There are twenty separate subsamples that make up the *DE-SynPUF*. Each subsample contains a random sample of the full *CMS Linkable 2008-2010 Medicare DE-SynPUF*. These twenty subsamples were created because the entire *DE-SynPUF* is too large to be distributed as a single file through the CMS web server.
- In each subsample, there are eight CSV files that contain the raw data for that subsample. The file name for each of the eight files contains the subsample number. Please see Table 1 for the specific file names.
- Each subsample contains all the beneficiary data and claims data for the subsample of beneficiaries. Users can work with anywhere from 1 to all 20 subsamples. The provided SAS READIN programs allow users to specify which subsamples to read in. These programs also include code that can concatenate as many of the subsamples as the user specifies as long as the subsamples are in consecutive numeric order.
- If users would like to use all of the data, they should download all twenty subsamples; if users would like to work with only one subsample, they can use any one of the 20 subsamples.
- If users would like to work with more than one subsample, but not all 20, they should download consecutively numbered subsamples to concatenate (e.g., subsample 1,2,3,4, and 5 but not subsamples 2, 4, 6, 8 and 10) as the SAS READIN programs assume the subsample numbers are sequential.
- If users would like to work with more than one non-consecutive subsample (e.g., subsamples 2, 4, 6, 8 and 10), users can write a DATA step to concatenate the files. Within each CSV file, the records are in order by *DESYNPUF_ID*. Therefore, if no modifications have been made to alter the order of the observations in the subsample SAS data sets, a BY *DESYNPUF_ID* statement in the DATA step that concatenates the subsamples will interleave the observations by the values of *DESYNPUF_ID*.

Downloading and Formatting the Data in SAS

1. Download ALL eight CSV files for each of the twenty subsamples of *DE-SynPUF* that users would like to work with from http://www.cms.gov/SynPUFs.

Table1. File Names of the Eight CSV Files Pertaining to Five File Types in each *DE-SynPUF* Subsample

File type	CSV File name	Number of Years of Data
Beneficiary Summary DE-SynPUF	DE1_0_2008_Beneficiary_Summary_File_Sample_#	1
	DE1_0_2009_Beneficiary_Summary_File_Sample_#	1
	DE1_0_2010_Beneficiary_Summary_File_Sample_#	1
Inpatient Claims DE-SynPUF	DE1_0_2008_to_2010_Inpatient_Claims_Sample_#	3
Outpatient Claims DE-SynPUF	DE1_0_2008_to_2010_Outpatient_Claims_Sample_#	3
Prescription Drug Events (PDE) DE-SynPUF	DE1_0_2008_to_2010_Prescription_Drug_Events_Sample_#	3
Carrier Claims DE-SynPUF	DE1_0_2008_to_2010_Carrier_Claims_Sample_#A	3
	DE1_0_2008_to_2010_Carrier_Claims_Sample_#B	3

NOTE: The "#" symbol takes on the values from 1 – 20 and is the subsample number (e.g., subsample 1 the *2008 Beneficiary Summary DE-SynPUF* is called "DE1_0_2008_Beneficiary_Summary_File_Sample_1")

2. Download the five SAS READIN* programs.

3. Unzip all files in the downloaded subsamples and put all unzipped files in the same folder on your server or hard drive. That is, no matter how many subsamples were downloaded, all unzipped CSV files should be put into the same folder.

4. Unzip the SAS READIN programs.

5. Open the SAS READIN program for the file type users would like to read into SAS (e.g., if users would like to read in Beneficiary Summary *DE-SynPUF*, users should use DESYNPUF_BENE_READIN.sas). Carefully read the instructions included in each of the five SAS READIN program before running any one of them.

6. After running the programs please validate the numbers of observations for each file against numbers provided in Table 2 on section 3 in the *CMS Linkable 2008-2010 Medicare DE-SynPUF* User Manual.

*NOTE: The SAS READIN programs read in CSV data files and transform them into SAS data sets. There are five SAS READIN programs: one for each file type. These READIN programs allow users to specify which subsamples to read in and to combine multiple subsamples if more than one subsample is downloaded and the subsamples are consecutively numbered.

Examples of Using SAS to Work with the *DE-SynPUF*

This section presents some helpful guidance on working with the *DE-SynPUF* in SAS. SAS commands are highlighted in **bold** while variables are highlighted in *italic*.

A 2008 *DE-SynPUF* beneficiary file contains data for synthetic beneficiaries enrolled in Medicare in 2008. A 2009 *DE-SynPUF* beneficiary file with the same suffix number contains data for the same beneficiaries who were alive and still enrolled in 2009. Similarly, a 2010 beneficiary file with the same suffix number contains data for the same beneficiaries who were alive and still enrolled in 2010.

<Case 1> Identifying Beneficiaries Enrolled in Different Time Periods

If you want to find the IDs for beneficiaries who enrolled in all three years, you can link the three data sets by the beneficiary ID variable *DESYNPUF_ID*. This DATA step will not save any variables from the beneficiary data sets. In this example, we use data only from subsample 1 (i.e., Suffix file name SAMPLE_1 indicates files from subsample 1).

```
DATA in080910;
   MERGE DE1_0_2008_BENE_SAMPLE_1(KEEP=desynpuf_id IN=in2008)
      DE1_0_2009_BENE_SAMPLE_1(KEEP=desynpuf_id IN=in2009)
      DE1_0_2010_BENE_SAMPLE_1(KEEP=desynpuf_id IN=in2010);
   BY desynpuf_id;

   IF in2008 AND in2009 AND in2010;

RUN;
```

<Case 2> Merging Beneficiary Data with Claims Data

If you want to merge beneficiary data for a specific year to a claims data set, you can link the two data sets by the beneficiary ID variable *DESYNPUF_ID*. Since the claims data sets contain data for all three years, if you want claims data for only the specific year, you will need to evaluate claims dates in the claims data set. In these examples, we use data only from subsample 1 (i.e., Suffix file name SAMPLE_1 indicates files from subsample 1).

<Scenario 1>

The following sample DATA step finds inpatient claims in 2008 for beneficiaries in the 2008 beneficiary file. The code evaluates *CLM_THRU_DT*, a variable in the inpatient file that specifies the claims through date. The observations in IP2008 are for beneficiaries who had at least one inpatient claim in 2008.

```
DATA ip2008;
  MERGE DE1_0_2008_BENE_SAMPLE_1(IN=inbene)
        DE1_0_2008_to_2010_IP_SAMPLE_1(IN=inip);
  BY desynpuf_id;

  IF inip AND inbene AND year(clm_thru_dt)=2008;

RUN;
```

<Scenario 2>

If you want to find beneficiaries who enrolled in all three years and had at least one inpatient claim from 2008 to 2010, adapt the following DATA step. The DATA step uses data set in080910 created in CASE 1 above.

```
DATA ip080910;
  MERGE in080910(IN= inall3)
    DE1_0_2008_to_2010_IP_SAMPLE_1(IN=inip);
  BY desynpuf_id;

  IF inip AND inall3;

RUN;
```

<Scenario 3>

If you want to find inpatient claims for beneficiaries who enrolled in all three years, had at least one inpatient claim from 2008 to 2010, and include some beneficiary variables in the output data set, adapt the following DATA step. This DATA step takes date of birth (*BENE_BIRTH_DT*), gender (*BENE_SEX_IDENT_CD*), and race (*BENE_RACE_CD*) information from the 2008 data set. It selects four chronic condition variables, *SP_ALZHDMTA, SP_CHF, SP_CHRNKIDN,* and *SP_CNCR* from the three beneficiary data sets. On the MERGE statement, these four chronic condition variables are renamed so that the information for each year is saved in the inpatient claim record.

```
DATA ip080910;
  MERGE in080910(IN=inall3)
        DE1_0_2008_BENE_SAMPLE_1(KEEP=desynpuf_id bene_birth_dt bene_sex_ident_cd
                                      bene_race_cd
                                      SP_ALZHDMTA SP_CHF SP_CHRNKIDN
                                      SP_CNCR
                                 RENAME=(SP_ALZHDMTA=SP_ALZHDMTA_08
                                      SP_CHF=SP_CHF_08
                                      SP_CHRNKIDN=SP_CHRNKIDN_08
                                      SP_CNCR=SP_CNCR_08))
        DE1_0_2009_BENE_SAMPLE_1(KEEP=desynpuf_id SP_ALZHDMTA SP_CHF
                                      SP_CHRNKIDN SP_CNCR
                                 RENAME=(SP_ALZHDMTA=SP_ALZHDMTA_09
                                      SP_CHF=SP_CHF_09
                                      SP_CHRNKIDN=SP_CHRNKIDN_09
                                      SP_CNCR=SP_CNCR_09))
        DE1_0_2010_BENE_SAMPLE_1(KEEP=desynpuf_id SP_ALZHDMTA SP_CHF
                                      SP_CHRNKIDN SP_CNCR
                                 RENAME=(SP_ALZHDMTA=SP_ALZHDMTA_10
                                      SP_CHF=SP_CHF_10
                                      SP_CHRNKIDN=SP_CHRNKIDN_10
                                      SP_CNCR=SP_CNCR_10))
        DE1_0_2008_to_2010_IP_SAMPLE_1(IN=inip);

  BY desynpuf_id;

  IF inall3 AND inip;

RUN;
```