USER MANUAL

Centers for Medicare and Medicaid Services (CMS) *Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)*

January 15, 2013

Table of Contents

List of Tables and Figures

1. *DE-SynPUF* Overview

The *CMS Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)* was designed to create a new type of file that would be useful for data entrepreneurs, for software and application development, and for research training purposes.

The *DE-SynPUF* contains multiple files per year for multiple years. The file contains synthesized data taken from a 5% random sample of Medicare beneficiaries in 2008 and their claims from 2008 to 2010. Each synthetic beneficiary was assigned a unique unidentifiable ID, *DESYNPUF_ID*, which is provided on each file to link synthetic claims to a synthetic beneficiary. This beneficiary ID carries no information about the enrollee or any patient records, and is provided solely for reference and data processing purposes.

Please find more details regarding how files of the *DE-SynPUF* were organized in Section 5 ("How to Link the *CMS Linkable 2008–2010 Medicare DE-SynPUF*") of this user manual. Files of the same type contain the same sets of variables for each year. Variable names in the *DE-SynPUF* were kept the same as those in the actual Medicare data unless they were significantly coarsened to decrease re-identification risk. In those cases, "SP_" was added to the original variable name for distinguishing.

Data Utility

The files preserve the detailed data structure and metadata of key variables at both the beneficiary and claim levels. However, the data are fully "synthetic," meaning no beneficiary in the *DE-SynPUF* is an actual Medicare beneficiary. They are all synthetic beneficiaries meant to represent actual beneficiaries. In order to protect the privacy of beneficiaries and to greatly reduce the risk of re-identification, a significant amount of interdependence and co-variation among variables has been altered in the synthetic process. The synthetic process used significantly diminishes the analytic utility of the file to produce reliable inferences about the actual Medicare beneficiary population (i.e., univariate statistics and regression coefficients produced with the *DE-SynPUF* will be biased).

Although the *DE-SynPUF* has limited empirical research utility, it does have the same data and file structure as the actual 5% Medicare beneficiary file and similar number of beneficiaries; it just has a smaller number of claims types and number of variables. Because the structure of the data is maintained, the *DE-SynPUF* is useful for building data tools that could be used with the actual data. Programs and procedures designed using the *DE-SynPUF* should be fully functional when applied to CMS Limited Data Sets (LDS) or Identifiable Data prior to 2011, assuming variable names have not been changed.

The *DE-SynPUF* can also be used to train researchers on how to conduct complex analyses of CMS claims data. Moreover, the *DE-SynPUF* will also provide a robust set of metadata on the CMS claims data that have not been available in the public domain.

The analytic utility of the data file differs based on the type and level of analysis being conducted:

- **Demographic**: The *DE-SynPUF* estimates of demographic characteristics (date of birth, date of death, sex, race, state, and county) of the beneficiary population match the univariate frequency of the full population of beneficiaries enrolled in Medicare at any time during the 2008 year.

- **Clinical**: The *DE-SynPUF* estimates for clinical variables such as chronic conditions can provide researchers with bounds on how many cases with a specific condition are likely to be in the Medicare claims, which could be used to generate power calculations for a grant application.

- **Economic/financial**: The *DE-SynPUF* estimates for the economic and financial variables provide a *lower bound* for the true estimate of cost for the full population of beneficiaries enrolled in Medicare at any time during the 2008 year and costs for 2009 and 2010 for this 2008 beneficiary example.

- **Multivariate modeling**: The dynamic relationships between variables (demographic, clinical, economic/financial, and provider information) were altered, to limit re-identification risk. Therefore, analyses from multivariate modeling should be interpreted with caution. However, the programs and procedures employed in the multivariate modeling will function on the CMS Limited Data Sets or Identifiable Data prior to 2011.

See Appendices A and B for more information.

2. Source Data for the *CMS Linkable 2008–2010 Medicare DE-SynPUF*

The source of the *CMS Beneficiary Summary DE-SynPUF* is a 5% sample of 2008 Medicare beneficiaries and their claims from 2008, 2009, and 2010. The sample used for the *CMS Linkable 2008–2010 DE-SynPUF* is disjoint from the existing 5% CMS research sample[1] and from existing Basic Stand Alone (BSA) Medicare Claims Public Use Files (PUFs).[2] Approximately 2.4 million beneficiaries in the source 5% sample of 2008 Medicare beneficiaries went through a stochastic de-identification process described in "Synthetic File Creation Process" (see Section 6) below. The *CMS 2008 Beneficiary Summary DE-SynPUF* contains synthesized Medicare beneficiaries. The *CMS 2008 Inpatient Claims DE-SynPUF,* the *CMS 2008 Outpatient Claims DE-SynPUF,* the *CMS 2008 Carrier Claims DE-SynPUF, and* the *CMS 2008 Prescription Drug Events (PDE) DE-SynPUF* are claims pertaining to those synthetic Medicare beneficiaries contained in the *CMS 2008 Beneficiary Summary DE-SynPUF*. The same applies to the 2009 and 2010 *DE-SynPUF*.

---

[1] Differences in How the Medicare 5% Files Are Generated. Technical Brief, ResDAC Publication Number TN-011, March 2007. Research Data Assistance Center, University of Minnesota, Minneapolis, MN.

[2] http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/index.html

3. Contents of the *CMS Linkable 2008–2010 Medicare DE-SynPUF*

The *CMS Linkable 2008–2010 Medicare DE-SynPUF* contains five types of files—the *CMS Beneficiary Summary DE-SynPUF,* the *CMS Inpatient Claims DE-SynPUF,* the *CMS Outpatient Claims DE-SynPUF,* the *CMS Carrier Claims DE-SynPUF,* and the *CMS Prescription Drug Events (PDE) DE-SynPUF—*from 2008 to 2010. Table 1 describes the number of variables, the unit of record, and the number of records available in the full *DE-SynPUF*. Table 2 provides the number of observations available in each of the data files for each of the *DE-SynPUF* subsamples.

Table 1.   Summary of the Characteristics of the *CMS Linkable 2008–2010 Medicare DE-SynPUF*

*DE-Syn*PUF	Unit of record	Number of records[*] 2008	Number of records[*] 2009	Number of records[*] 2010
Beneficiary Summary DE-SynPUF	Beneficiary	2,326,856	2,291,320	2,255,098
Inpatient Claims DE-SynPUF [*]	Claim	547,800	504,941	280,081
Outpatient Claims DE-SynPUF [*]	Claim	5,673,808	6,519,340	3,633,839
Carrier Claims DE-SynPUF [*]	Claim	34,276,324	37,304,993	23,282,135
Prescription Drug Events (PDE) DE-SynPUF [*]	Claim	39,927,827	43,379,293	27,778,849

Note: Claim counts for 2010 are lower because of attrition from death and some effects of disclosure treatment.
* Year is determined by the claim through date (*CLM_THRU_DT*)

Table 2. Observation Numbers by the *DE-SynPUF* Type in Each Subsample

Subsample	Beneficiary 2008	Beneficiary 2009	Beneficiary 2010	Carrier* A+B#	Carrier* A#	Carrier* B#	IP*	OP*	PDE*
1	116,352	114,538	112,754	4,741,335	2,370,667	2,370,668	66,773	790,790	5,552,421
2	116,395	114,618	112,845	4,745,914	2,372,957	2,372,957	66,494	792,562	5,561,154
3	116,390	114,644	112,812	4,745,604	2,372,802	2,372,802	66,672	792,415	5,557,147
4	116,279	114,528	112,699	4,743,079	2,371,539	2,371,540	66,253	789,485	5,549,070
5	116,364	114,539	112,687	4,735,118	2,367,559	2,367,559	66,414	790,538	5,549,634
6	116,234	114,532	112,713	4,745,684	2,372,842	2,372,842	66,977	793,146	5,557,441
7	116,352	114,569	112,747	4,741,798	2,370,899	2,370,899	66,791	791,916	5,560,085
8	116,330	114,589	112,777	4,738,432	2,369,216	2,369,216	66,490	790,244	5,556,025
9	116,287	114,495	112,685	4,734,533	2,367,266	2,367,267	66,763	790,818	5,552,470
10	116,353	114,588	112,769	4,747,365	2,373,682	2,373,683	66,585	791,355	5,545,284
11	116,339	114,459	112,586	4,743,674	2,371,837	2,371,837	66,425	791,845	5,552,888
12	116,387	114,646	112,886	4,745,585	2,372,792	2,372,793	66,717	789,209	5,555,572
13	116,363	114,511	112,665	4,742,836	2,371,418	2,371,418	66,324	790,950	5,549,202
14	116,374	114,637	112,786	4,744,374	2,372,187	2,372,187	67,024	792,115	5,553,031
15	116,391	114,612	112,809	4,737,282	2,368,641	2,368,641	66,846	791,462	5,551,438
16	116,377	114,626	112,901	4,744,582	2,372,291	2,372,291	66,800	792,099	5,560,073
17	116,285	114,422	112,615	4,745,896	2,372,948	2,372,948	66,495	790,146	5,555,828
18	116,405	114,638	112,876	4,745,177	2,372,588	2,372,589	66,428	793,115	5,564,559
19	116,224	114,488	112,675	4,747,966	2,373,983	2,373,983	67,037	792,733	5,550,500
20	116,375	114,641	112,811	4,747,218	2,373,609	2,373,609	66,514	790,044	5,552,147

NOTE: # Because of file size limitations, each carrier subsample was split into two CSV files as segment A and segment B.
IP: Inpatient
OP: Outpatient
PDE: Prescription Drug Events
* Three years of data, from 2008–2010

4. Summary of Variables of the *CMS Linkable 2008–2010 Medicare DE-SynPUF*

Please refer to the *CMS Linkable 2008–2010 Medicare DE-SynPUF Codebook* for more detailed information about the data. Next, we list the variables and their labels in each of the files.

1. The *CMS Beneficiary Summary DE-SynPUF* contains 32 variables. Each record pertains to a

    synthetic Medicare beneficiary and contains:

#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*BENE_BIRTH_DT*	DESYNPUF: Date of birth
3	*BENE_DEATH_DT*	DESYNPUF: Date of death
4	*BENE_SEX_IDENT_CD*	DESYNPUF: Sex
5	*BENE_RACE_CD*	DESYNPUF: Beneficiary Race Code
6	*BENE_ESRD_IND*	DESYNPUF: End stage renal disease Indicator
7	*SP_STATE_CODE*	DESYNPUF: State Code
8	*BENE_COUNTY_CD*	DESYNPUF: County Code
9	*BENE_HI_CVRAGE_TOT_MONS*	DESYNPUF: Total number of months of part A coverage for the beneficiary.
10	*BENE_SMI_CVRAGE_TOT_MONS*	DESYNPUF: Total number of months of part B coverage for the beneficiary.
11	*BENE_HMO_CVRAGE_TOT_MONS*	DESYNPUF: Total number of months of HMO coverage for the beneficiary.
12	*PLAN_CVRG_MOS_NUM*	DESYNPUF: Total number of months of part D plan coverage for the beneficiary.
13	*SP_ALZHDMTA*	DESYNPUF: Chronic Condition: Alzheimer or related disorders or senile
14	*SP_CHF*	DESYNPUF: Chronic Condition: Heart Failure
15	*SP_CHRNKIDN*	DESYNPUF: Chronic Condition: Chronic Kidney Disease
16	*SP_CNCR*	DESYNPUF: Chronic Condition: Cancer
17	*SP_COPD*	DESYNPUF: Chronic Condition: Chronic Obstructive Pulmonary Disease
18	*SP_DEPRESSN*	DESYNPUF: Chronic Condition: Depression
19	*SP_DIABETES*	DESYNPUF: Chronic Condition: Diabetes
20	*SP_ISCHMCHT*	DESYNPUF: Chronic Condition: Ischemic Heart Disease
21	*SP_OSTEOPRS*	DESYNPUF: Chronic Condition: Osteoporosis
22	*SP_RA_OA*	DESYNPUF: Chronic Condition: rheumatoid arthritis and osteoarthritis (RA/OA)
23	*SP_STRKETIA*	DESYNPUF: Chronic Condition: Stroke/transient Ischemic Attack
24	*MEDREIMB_IP*	DESYNPUF: Inpatient annual Medicare reimbursement amount
25	*BENRES_IP*	DESYNPUF: Inpatient annual beneficiary responsibility amount
26	*PPPYMT_IP*	DESYNPUF: Inpatient annual primary payer reimbursement amount

#	Variable names	Labels
27	*MEDREIMB_OP*	DESYNPUF: Outpatient Institutional annual Medicare reimbursement amount
28	*BENRES_OP*	DESYNPUF: Outpatient Institutional annual beneficiary responsibility amount
29	*PPPYMT_OP*	DESYNPUF: Outpatient Institutional annual primary payer reimbursement amount
30	*MEDREIMB_CAR*	DESYNPUF: Carrier annual Medicare reimbursement amount
31	*BENRES_CAR*	DESYNPUF: Carrier annual beneficiary responsibility amount
32	*PPPYMT_CAR*	DESYNPUF: Carrier annual primary payer reimbursement amount

2. The *CMS Inpatient Claims DE-SynPUF* contains 81 variables. Each record pertains to a synthetic inpatient claim and contains:

#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*CLM_ID*	DESYNPUF: Claim ID
3	*SEGMENT*	DESYNPUF: Claim Line Segment
4	*CLM_FROM_DT*	DESYNPUF: Claims start date
5	*CLM_THRU_DT*	DESYNPUF: Claims end date
6	*PRVDR_NUM*	DESYNPUF: Provider Institution
7	*CLM_PMT_AMT*	DESYNPUF: Claim Payment Amount
8	*NCH_PRMRY_PYR_CLM_PD_AMT*	DESYNPUF: NCH Primary Payer Claim Paid Amount
9	*AT_PHYSN_NPI*	DESYNPUF: Attending Physician – National Provider Identifier Number
10	*OP_PHYSN_NPI*	DESYNPUF: Operating Physician – National Provider Identifier Number
11	*OT_PHYSN_NPI*	DESYNPUF: Other Physician – National Provider Identifier Number
12	*CLM_ADMSN_DT*	DESYNPUF: Inpatient admission date
13	*ADMTNG_ICD9_DGNS_CD*	DESYNPUF: Claim Admitting Diagnosis Code
14	*CLM_PASS_THRU_PER_DIEM_AMT*	DESYNPUF: Claim Pass Thru Per Diem Amount
15	*NCH_BENE_IP_DDCTBL_AMT*	DESYNPUF: NCH Beneficiary Inpatient Deductible Amount
16	*NCH_BENE_PTA_COINSRNC_LBLTY_AM*	DESYNPUF: NCH Beneficiary Part A Coinsurance Liability Amount
17	*NCH_BENE_BLOOD_DDCTBL_LBLTY_AM*	DESYNPUF: NCH Beneficiary Blood Deductible Liability Amount
18	*CLM_UTLZTN_DAY_CNT*	DESYNPUF: Claim Utilization Day Count
19	*NCH_BENE_DSCHRG_DT*	DESYNPUF: Inpatient discharged date
20	*CLM_DRG_CD*	DESYNPUF: Claim Diagnosis Related Group Code
21-30	*ICD9_DGNS_CD_1 – ICD9_DGNS_CD_10*	DESYNPUF: Claim Diagnosis Code 1 – Claim Diagnosis Code 10
31-36	*ICD9_PRCDR_CD_1 – ICD9_PRCDR_CD_6*	DESYNPUF: Claim Procedure Code 1 – Claim Procedure Code 6
37-81	*HCPCS_CD_1 – HCPCS_CD_45*	DESYNPUF: Revenue Center HCFA Common Procedure Coding System 1 – Revenue Center HCFA Common Procedure Coding System 45

3. The *CMS Outpatient Claims DE-SynPUF* contains 76 variables. Each record pertains to a synthetic outpatient claim and contains:

#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*CLM_ID*	DESYNPUF: Claim ID
3	*SEGMENT*	DESYNPUF: Claim Line Segment
4	*CLM_FROM_DT*	DESYNPUF: Claims start date
5	*CLM_THRU_DT*	DESYNPUF: Claims end date
6	*PRVDR_NUM*	DESYNPUF: Provider Institution
7	*CLM_PMT_AMT*	DESYNPUF: Claim Payment Amount
8	*NCH_PRMRY_PYR_CLM_PD_AMT*	DESYNPUF: NCH Primary Payer Claim Paid Amount
9	*AT_PHYSN_NPI*	DESYNPUF: Attending Physician – National Provider Identifier Number
10	*OP_PHYSN_NPI*	DESYNPUF: Operating Physician – National Provider Identifier Number
11	*OT_PHYSN_NPI*	DESYNPUF: Other Physician – National Provider Identifier Number
12	*NCH_BENE_BLOOD_DDCTBL_LBLTY_AM*	DESYNPUF: NCH Beneficiary Blood Deductible Liability Amount
13-22	*ICD9_DGNS_CD_1 – ICD9_DGNS_CD_10*	DESYNPUF: Claim Diagnosis Code 1 – Claim Diagnosis Code 10
23-28	*ICD9_PRCDR_CD_1 – ICD9_PRCDR_CD_6*	DESYNPUF: Claim Procedure Code 1 – Claim Procedure Code 6
29	*NCH_BENE_PTB_DDCTBL_AMT*	DESYNPUF: NCH Beneficiary Part B Deductible Amount
30	*NCH_BENE_PTB_COINSRNC_AMT*	DESYNPUF: NCH Beneficiary Part B Coinsurance Amount
31	*ADMTNG_ICD9_DGNS_CD*	DESYNPUF: Claim Admitting Diagnosis Code
32-76	*HCPCS_CD_1 – HCPCS_CD_45*	DESYNPUF: Revenue Center HCFA Common Procedure Coding System 1 – Revenue Center HCFA Common Procedure Coding System 45

4. The *CMS Carrier Claims DE-SynPUF* contains 142 variables. Each record pertains to a synthetic physician/supplier claim and contains:

#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*CLM_ID*	DESYNPUF: Claim ID
3	*CLM_FROM_DT*	DESYNPUF: Claims start date
4	*CLM_THRU_DT*	DESYNPUF: Claims end date
5-12	*ICD9_DGNS_CD_1 – ICD9_DGNS_CD_8*	DESYNPUF: Claim Diagnosis Code 1 – Claim Diagnosis Code 8
13-25	*PRF_PHYSN_NPI_1 – PRF_PHYSN_NPI_13*	DESYNPUF: Provider Physician – National Provider Identifier Number
26-38	*TAX_NUM_1 – TAX_NUM_13*	DESYNPUF: Provider Institution Tax Number
39-51	*HCPCS_CD_1 – HCPCS_CD_13*	DESYNPUF: Line HCFA Common Procedure Coding System 1 – Line HCFA Common Procedure Coding System 13
52-64	*LINE_NCH_PMT_AMT_1– LINE_NCH_PMT_AMT_13*	DESYNPUF: Line NCH Payment Amount 1 – Line NCH Payment Amount 13
65-77	*LINE_BENE_PTB_DDCTBL_AMT_1 – LINE_BENE_PTB_DDCTBL_AMT_13*	DESYNPUF: Line Beneficiary Part B Deductible Amount 1 – Line Beneficiary Part B Deductible Amount 13
78-90	*LINE_BENE_PRMRY_PYR_PD_AMT_1 – LINE_BENE_PRMRY_PYR_PD_AMT_13*	DESYNPUF: Line Beneficiary Primary Payer Paid Amount 1 – Line Beneficiary Primary Payer Paid Amount 13
91-103	*LINE_COINSRNC_AMT_1 – LINE_COINSRNC_AMT_13*	DESYNPUF: Line Coinsurance Amount 1 – Line Coinsurance Amount 13
104-116	*LINE_ALOWD_CHRG_AMT_1 – LINE_ALOWD_CHRG_AMT_13*	DESYNPUF: Line Allowed Charge Amount 1 – Line Allowed Charge Amount 13
117-129	*LINE_PRCSG_IND_CD_1 – LINE_PRCSG_IND_CD_13*	DESYNPUF: Line Processing Indicator Code 1 – Line Processing Indicator Code13
130-142	*LINE_ICD9_DGNS_CD_1 – LINE_ICD9_DGNS_CD_13*	DESYNPUF: Line Diagnosis Code 1 – Line Diagnosis Code 13

5. The *CMS Prescription Drug Events (PDE) DE-SynPUF* contains 8 variables. Each record pertains to a synthetic Part D event and contains:

#	Variable names	Labels
1	*DESYNPUF_ID*	DESYNPUF: Beneficiary Code
2	*PDE_ID*	DESYNPUF: CCW Part D Event Number
3	*SRVC_DT*	DESYNPUF: RX Service Date
4	*PROD_SRVC_ID*	DESYNPUF: Product Service ID
5	*QTY_DSPNSD_NUM*	DESYNPUF: Quantity Dispensed
6	*DAYS_SUPLY_NUM*	DESYNPUF: Days Supply
7	*PTNT_PAY_AMT*	DESYNPUF: Patient Pay Amount
8	*TOT_RX_CST_AMT*	DESYNPUF: Gross Drug Cost

5. How to Link the *CMS Linkable 2008–2010 Medicare DE-SynPUF*

The unique identifier *DESYNPUF_ID* was specifically created for the *DE-SynPUF* to identify synthetic beneficiaries. This identifier carries no information about the patient or any patient records, and is provided solely for reference and data processing purposes.

Twenty random subsamples of each file type included in the *DE-SynPUF* were created because the entire *DE-SynPUF* file is too large to be distributed as a single downloadable file. This design allows *DE-SynPUF* users who do not need the entire synthetic population of the *DE-SynPUF* to download only as many subsamples as they desire (between 1 and 20). One reason for downloading only one subsample is that working with even one of these subsamples on a personal computer can require a significant amount of computing resources and be time intensive.

Each subsample contains eight CSV files: three beneficiary files (one for each year), one inpatient file containing three years of data, one outpatient file containing three years of data, one PDE file containing three years of data, and two carrier files containing three years of data (segment A and segment B). Because beneficiary files contain time-varying variables such as chronic conditions, reimbursement, and death, three files (one for each year) were provided to keep the variable names the same as the ones in the actual data. If the beneficiary died between 2008 and 2010, there will not be claims data for that beneficiary in the years after death. Because of file size limitations, each carrier subsample was split into two CSV files as segment A and segment B, both of which must be downloaded. The *DE-SynPUF* provides three years of claims in one file to decrease the number of files users have to download and to take the advantage of multiple years of data. Table 3 shows file names of the eight CSV Files pertaining to the five file types in Each *DE-SynPUF* subsample.

Figure 1 shows an example of how a beneficiary file with five beneficiaries is linked with two claims-based files: one inpatient files with three claims and one outpatient file with four claims. The files are linked through the unique identifier DESYNPUF_ID. In the figure, one of the beneficiaries has one inpatient claim but no outpatient claims; another beneficiary has no inpatient claims but three outpatient claims; a third beneficiary has two inpatient claims and one outpatient claim; two of the beneficiaries have no inpatient or outpatient claims. Because of the nature of claims data, not all beneficiaries have all types of claims. That is, some beneficiaries have multiple claims in multiple claim-based data files; some may have a single claim in a single claims-based data file; and some have no claims in any of the claims-based data files.

Table 3. File Names of the Eight CSV Files Pertaining to the 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_# DE1_0_2009_Beneficiary_Summary_File_Sample_# 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 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 of the 2008 Beneficiary Summary *DE-SynPUF* is called "DE1_0_2008_Beneficiary_Summary_File_Sample_1")

Figure 1. Claims Data File Structure
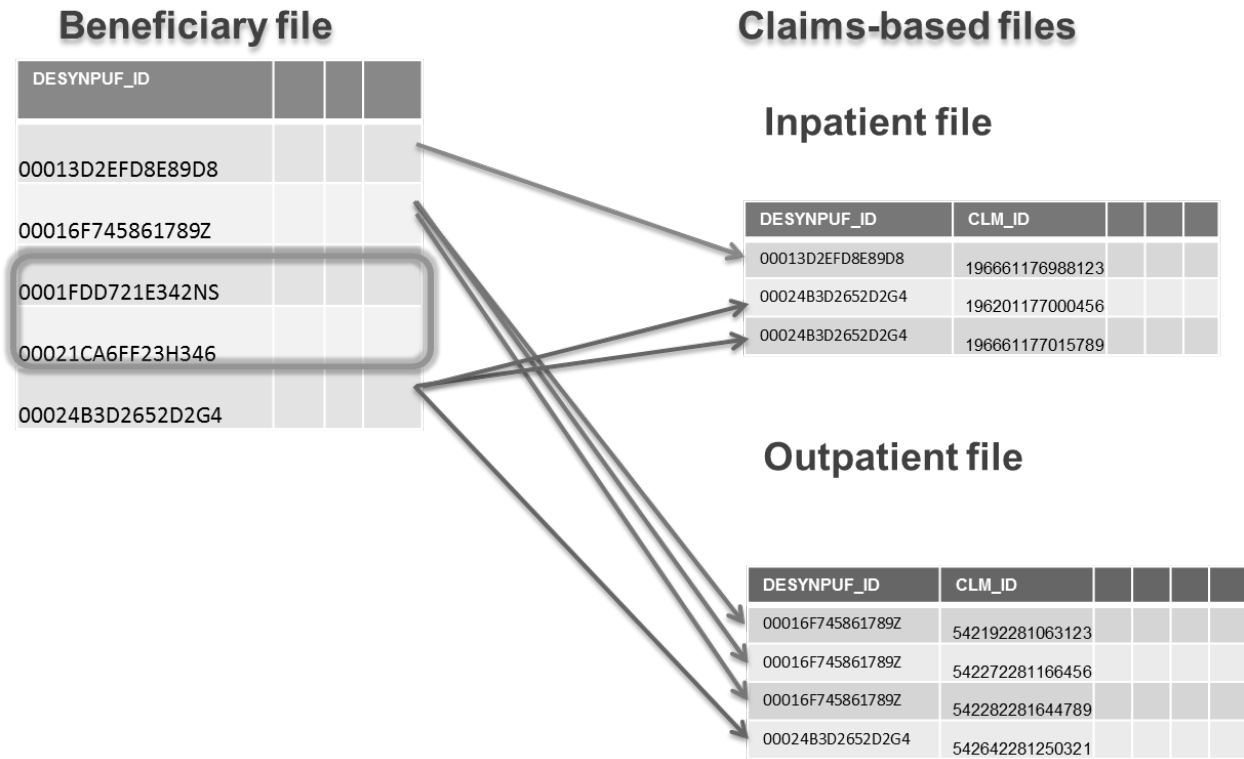
6. Methodology and Limitation

Synthetic File Creation Process

The *DE-SynPUF* was created by starting with an actual beneficiary as a "seed" for a synthetic beneficiary. The variables of the seed beneficiary profile were changed by taking characteristics from similar but different "donor" beneficiaries within the source data. The claims from the seed beneficiary were then replaced with claims from other donor beneficiary claims sets. There were processes in place to ensure that a synthetic claim set was created from at least three different donors. The claims from the seed beneficiary were further altered by changing variable values (e.g., coarsening, imputation) within the claim that could have caused the claim to disclose information about the donor beneficiary. Many of these synthetic processes were designed to reflect trends observed in the data while protecting against disclosing information about the actual beneficiaries used in the creation of the synthetic beneficiaries.

A variety of statistical disclosure limitation techniques were used to protect the confidentiality of beneficiary data in the *CMS Linkable 2008–2010 Medicare DE-SynPUF*. Synthetic beneficiaries and their claims were based on actual seed beneficiaries. However, disclosure was reduced through multiple deterministically or stochastically applied treatment methods. First, hot decking-based procedures (Methods used to find similar records in a data set, often for the purpose of imputing missing values.) were used to find donors for beneficiary-level variables and individual claims. Second, other synthetic processes described below were used to protect other elements of the data. A key feature was that each synthetic record is based on information from multiple donors and each donor can contribute only a very small "chunk" of data. The disclosure limitation methods used in the creation of the *DE-SynPUF* were:

- **Variable Reduction**: The number of variables in each table was limited to a set that was useful and appropriate for development users.

- **Suppression**: Records, whether beneficiary or claim, that were rare in the data and that had disclosure risk even in a synthetic file were removed by applying appropriate k-anonymity rules based on either population- or sample-specific counts.

- **Substitution**: Variable values were altered by replacing them with values from a similar donor record based on key variables (i.e., conditional on matching certain variables).

- **Imputation**: Values of single variables were altered by drawing values from empirical distributions conditioned on key variables. The empirical distributions were first coarsened and truncated, removing potentially identifying values.

- **Date Perturbation**: Timelines were altered by changing dates and intervals between events.

- **Coarsening**: Continuous variables (e.g., year of birth or expenditures) were coarsened enough to limit disclosure but remain realistically useful.

The disclosure limitation methods applied at the beneficiary level were:

- Reducing the number of variables from 240 to 32.

- Completely suppressing rare or extreme beneficiaries from the data.

- Substituting geo-demographic values from donor beneficiaries or randomly from empirical distributions (e.g., county codes).

- Synthesizing chronic condition indicators using synthetic claims created by other processes.

The disclosure limitation methods applied at claim level were:

- Drastically reducing the number of variables in a claim record—for example, 81 out of 611 variables were kept for inpatient claims.

- Suppressing rare diagnosis and procedure codes from the data used to create synthetic claims.

- Substituting claims from donor beneficiaries using an actual beneficiary as the seed, or pattern, for the synthetic beneficiaries; donor claims were found using a key variable from the seed and donor claims.

- Restricting the amount of information coming from any one donor and always using multiple donors; a minimum of three donors contributed to each single synthetic beneficiary claim set.

- Synthesizing secondary variable sets within the donated claims conditioned on key variables, for added disclosure protection.

- Perturbing various claim dates by altering the start date of the claim set used as the seed and proportionally altering the number of days between claims.

- Coarsening expenditure variables so that larger values were coarsened into larger bins and truncating both tails of the distribution (top and bottom).

- Synthesizing provider information (institution and physician) by drawing from empirical distribution conditioned on the synthesized geography of the beneficiary.

- Suppressing rare combinations of institution and physician codes from the data used to create synthetic claims.

In addition, to improve provider confidentiality, some of the disclosure limitation methods described above were implemented on provider variables. All codes identifying provider institutions or physicians were altered to reduce the likelihood of their identification. Synthetic providers and physicians are associated with synthetic beneficiaries based solely on the geography of the synthetic beneficiaries they serve. Because these fields are random numbers/characters, with no association to any known id number, analytic inferences to the Medicare providers should not be made when using these variables.

Limitations of the *CMS Linkable 2008–2010 Medicare DE-SynPUF*

Because the *DE-SynPUF* has the structure of Medicare data, these synthetic files suffer the same data limitations as do Medicare data. The Research Data Assistance Center (ResDAC) offers *ResConnect* to provide researchers with an in-depth explanation of common CMS procedures, files, variables, and utilities.[3] The Chronic Condition Data Warehouse Web site also offers analytic guidance.[4] Users of the file who are not familiar with Medicare data are strongly advised to visit ResDAC or the Chronic Condition Data Warehouse to learn more about Medicare data whenever they have questions about using the *DE-SynPUF*.

---

[3] Research Data Assistance Center. ResConnect. http://www.resdac.org/resconnect

[4] Chronic Condition Data Warehouse. "Analytic Guidance." http://www.ccwdata.org/analytic-guidance/index.htm

Moreover, all variables in the *DE-SynPUF* are imputed, suppressed, and coarsened as part of disclosure treatment. As a result, the *DE-SynPUF* has very limited inferential research utility because of the synthetic process used to generate the data. That is, analyses using the *DE-SynPUF* to draw inferences about Medicare beneficiaries, providers, or the Medicare program will be misleading and often incorrect.

7. Support for the *DE-SynPUF* and Further Information

Questions and comments about the *CMS Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use Files (DE-SynPUF)* can be submitted to ResDAC[5] at the University of Minnesota at 1-888-9RESDAC or by e-mail at resdac@umn.edu.

---

[5] http://www.resdac.org/

Appendix A: Comparison of the *DE-SynPUF* with Actual Data

The *DE-SynPUF* preserves the detailed data structure of key variables at both the beneficiary and claim levels. However, the data are fully synthetic for disclosure safety. As a result, much of the interdependence and co-variation among variables has been altered on purpose to reduce the risk of re-identification, which significantly diminishes the true analytic utility of the file to produce reliable univariate and multivariate statistics such as conditional correlations and regression coefficients. Here we provide some estimates to show the differences between statistics calculated using the *DE-SynPUF* and statistics calculated on an actual 5% sample of Medicare beneficiaries.

These results are not meant to provide support for the idea of using the *DE-SynPUF* to draw inferences about the Medicare population. They are simply reported as a tool for users to better understand the limits of the *DE-SynPUF*. These analyses should **not** be interpreted as providing a guide to making a statistical adjustment (or as providing bounds) when using the *DE-SynPUF* to draw inferences about the Medicare beneficiary population. Any specific analysis users would conduct with the *DE-SynPUF* (and compare with an actual sample of Medicare beneficiaries and their claims) could produce very different results (either worse or better) from those presented in this section.

I. Number of Claims per Beneficiary by Service Type Over Three Years

Table 4. Comparison of Estimates from the *DE-SynPUF* and an Actual Medicare 5% Beneficiary Sample by Claim Types—Distribution of Number of Claims per Beneficiary over Three Years

Claim Type	Types	10%	20%	80%	90%
IP	*DE-SynPUF*	1	1	3	4
IP	Actual	1	1	4	5
OP	*DE-SynPUF*	2	3	16	21
OP	Actual	2	3	21	34
CAR	*DE-SynPUF*	4	12	99	104
CAR	Actual	5	15	103	147
PDE	*DE-SynPUF*	3	5	103	137
PDE	Actual	14	30	174	242

NOTE:
IP: Inpatient
OP: Outpatient
CAR: Carrier
PDE: Prescription Drug Events

II. Demography

Table 5. Comparison of Estimates from the *DE-SynPUF* with an Actual 2008 Medicare 5% Beneficiary Sample—Demography

	2008 *DE-SynPUF (%)*	**2008 (%)[1]**
Total	**2.3 million**	**2.4 million**
Sex		
Male	44.4	45
Female	55.6	55
Race/Ethnicity		
White	82.8	83
Black	10.7	10
Other/Hispanic	6.6	6
Year of Birth		
Pre-1924	12.8	13
1924–1928	12.5	12
1929–1933	15.6	15
1934–1938	18.9	19
1939–1943	23.8	24
Post-1943	16.3	16

Note: [1] Rounded to the nearest whole numbers. Percentages may not add up to 100% due to editing to prevent disclosure.

III. Claimants by Service Type by Year

Table 6. Comparison of Estimates from the *DE-SynPUF* with an Actual 2008 Medicare 5% Beneficiary Sample—Claimants by Service Type in 2008, 2009, and 2010

	2008 *DE-SynPUF* **Percent (%)[1]**	**2008** **Percent (%)[1,2]**	**2009** *DE-SynPUF* **Percent (%)[1]**	**2009** **Percent (%)[1,2]**	**2010** *DE-SynPUF* **Percent (%)[1]**	**2010** **Percent (%)[1,2]**
Inpatient	13.6	16	16.4	15	10.7	15
Outpatient	51.3	50	62.9	50	49.2	50
Carrier	73.4	70	80.0	70	76.0	70
PDE	62.6	53	79.0	56	73.5	57

NOTE: [1] Percent of beneficiaries with at least one claim in a certain claim type.
[2] Rounded to the nearest whole numbers. Percentages may not add up to 100% due to editing to prevent disclosure.
PDE: Prescription Drug Events

IV. Numbers of Claims of Claimants by Service Type by Year

Table 7. Comparison of Estimates from the *DE-SynPUF* with an Actual 2008 Medicare 5% Beneficiary Sample—Mean Number of Claims for Claimants by Service Type in 2008, 2009, and 2010

	2008 *DE-SynPUF* Mean	2008 Mean[1]	2009 *DE-SynPUF* Mean	2009 Mean[1]	2010 *DE-SynPUF* Mean	2010 Mean[1]
Inpatient	1.7	2	1.3	2	1.2	2
Outpatient	4.8	6	4.5	6	3.3	6
Carrier	20.1	25	20.4	26	13.6	26
PDE	27.4	41	24.0	42	16.8	42

NOTE: [1] Rounded to the nearest whole numbers. Percentages may not add up to 100% due to editing to prevent disclosure.
PDE: Prescription Drug Events

V. Reimbursement by Source by Year

Table 8. Comparison of Estimates from the *DE-SynPUF* with an Actual 2008 Medicare 5% Beneficiary Sample—Reimbursement by Source in 2008, 2009, and 2010

	2008 *DE-SynPUF* Mean	2008 Mean[1]	2009 *DE-SynPUF* Mean	2009 Mean[1]	2010 *DE-SynPUF* Mean	2010 Mean[1]
Inpatient						
Total	$2,544	$2,900	$2,519	$3,000	$1,441	$3,100
Medicare paid	$2,194	$2,500	$2,177	$2,700	$1,244	$2,700
Beneficiary paid	$247	$200	$248	$200	$145	$200
3rd party paid	$103	$100	$94	$100	$52	$100
Outpatient						
Total	$846	$1,100	$1,028	$1,200	$580	$1,300
Medicare paid	$624	$800	$765	$900	$434	$1,000
Beneficiary paid	$197	$300	$234	$300	$131	$300
3rd party paid	$25	*	$29	*	$15	*
Carrier						
Total	$1,536	$2,100	$1,734	$2,300	$1,100	$2,400
Medicare paid	$1,172	$1,600	$1,338	$1,800	$848	$1,800
Beneficiary paid	$346	$500	$375	$500	$239	$500
3rd party paid	$19	*	$21	*	$13	*
PDE[2]						
Total	$1,965	$3,200	$1,725	$3,300	$1,192	$3,400
Medicare paid	$55	$100	$56	$100	$57	$100
Beneficiary paid	$10	*	$10	*	$10	*

NOTE: [1] Rounded to the nearest multiple of $100.
[2] Data on 3rd party paid is not available in PDE - Prescription Drug Events.

VI.Chronic Conditions

In this section, we compare chronic condition estimates created from the *DE-SynPUF* with chronic condition estimates created from an actual 5% sample of 2008 Medicare beneficiaries. We calculate the estimates from the actual 5% sample of Medicare beneficiaries for two groups and the results are presented in Table 9. The two groups are:

- **Actual Estimate**: Based on an full actual 5% sample and including beneficiaries who had no claims in a given year

- **Actual Claimant Estimate**: Restricted to beneficiaries with at least one claim in a given year

In Table 9 most prevalence rates of chronic conditions in the *DE-SynPUF* are higher than those in the actual 2008 Medicare 5% beneficiary sample. The *DE-SynPUF* chronic condition estimates tend to be closer to (although still higher than) those for the claimant sub-population due to the synthetic processes used to construct the *DE-SynPUF*. These analyses aim to show some selected differences between the *DE-SynPUF* and an actual 5% sample and should **not** be interpreted as providing a guide to making a statistical adjustment (or as providing bounds) when using the *DE-SynPUF* to draw inferences about the Medicare beneficiary population. Any specific analysis users would want to do with the *DE-SynPUF* could produce very different from the results in Table 9 (either closer to the actual estimate or further from the actual estimate).

Table 9. Comparison of Estimates from the *DE-SynPUF* with an Actual 2008 Medicare 5% Beneficiary Sample—Chronic Conditions in 2008

Disease	Actual Estimate[1] (%)	Actual Claimant Estimate[1] (%)	DE-SynPUF (%)
Glaucoma	7	8	7.0
Chronic Obstructive Pulmonary Disease	8	9	13.5
Alzheimer's and related disorders or senile	8	9	19.2
Osteoporosis	9	10	17.4
Kidney Disease	9	11	16.1
Depression	9	11	21.3
Heart Failure	12	14	28.5
Cataract(s)	14	17	18.3
Rheumatoid Arthritis/ Osteoarthritis	15	17	15.4
Diabetes	20	23	38.0
Ischemic Heart Disease	23	27	42.0

NOTE: [1] Rounded to the nearest whole numbers due to editing to prevent disclosure.
[2] Similar results were found in 2009 and 2010.

Appendix B: Examples of the *DE-SynPUF* Application

To further explore the application of the *DE-SynPUF*, the following section presents a comparative effectiveness research (CER) example and a health services research (HSR) example that compared the analysis completed on the *DE-SynPUF* with the same analysis completed on an actual 5% sample of Medicare beneficiaries and their claims.

Both examples are solely intended to show that the *DE-SynPUF* preserves the detailed data file structure at both the beneficiary and claim levels, so it allows data entrepreneurs to develop and create software and applications that can be applied to actual CMS claims data or to train researchers to use CMS claims data. The *DE-SynPUF* has been designed so that programs and procedures created within it will function on the CMS Limited Data Sets or Identifiable Data prior to 2011.

These results are not meant to provide support for the idea of using the *DE-SynPUF* to draw inferences about the Medicare population. They are simply reported as a tool for users to better understand the limits of the *DE-SynPUF*. These analyses should **not** be interpreted as providing a guide to making a statistical adjustment (or as providing bounds) when using the *DE-SynPUF* to draw inferences about the Medicare beneficiary population. Any specific analysis users would conduct with the *DE-SynPUF* (and compare with an actual sample of Medicare beneficiaries and their claims) could produce very different results (either worse or better) from those presented in this section.

I. A CER Example

For this basic CER example, we were interested in the outcomes of angioplasty (*aka* percutaneous coronary interventions, denoted PTCA[6] hereafter) versus coronary artery bypass grafting (denoted CABG[7] hereafter). We used variables including Claim Diagnosis Code, Claim Procedure Code, Revenue Center HCFA Common Procedure Coding System (HCPCS), and Claim Diagnosis Related Group Code (DRG) to identify procedures or diseases of interest. We also used variable Claims Dates to decide the sequence of events. In this example, we used 2008-2010 claims of the *DE-SynPUF* and actual claims for a 5% sample of Medicare beneficiaries. To complete the analysis, we pulled data from the beneficiary summary, inpatient, outpatient, and carrier files of both the *DE-SynPUF* and the 5% Medicare beneficiary sample.

We studied a cohort of patients who had PTCA or CABG in an inpatient setting in 2009. The first hospitalization event with either PTCA or CABG performed was defined as the index event. The admission date of this index event was defined as index admission date while the discharge date of this index event was defined as index discharge date. Outcomes are defined as specific codes found in claims *after* the interventions, while comorbidities are defined as specific codes found in claims *before* the interventions. Therefore, the outcome of interest in this study was any occurrence of cardiac events including angina, heart failure, ischemic heart disease, or cardiac arrest during the 12 months after the index discharge date. Comorbidities were defined as any occurrence of hypertension, high cholesterol, diabetes, ischemic heart disease during the 12 months before the index admission date. An occurrence was defined as one of the special codes listed in Table 10 found in the inpatient, outpatient, or carrier files. Because we defined index events using 2009 data, we also utilized 2008 and 2010 data to find any occurrence of comorbidities or outcomes as defined above.

We ran a multiple variable logistic regression on outcome—any heart failure, for example, within 12 months after discharge from the PTCA or CABG procedure. Demographics (sex, age, and race/ethnicity) and comorbidities (hypertension, high cholesterol, diabetes, and previous ischemic heart disease) were included as covariates. Table 11 shows the results of bivariate analysis of intervention (PTCA or CABG) and demographics, comorbidities, and outcomes. Table 12 presents results from a logistic regression with heart failure as an outcome of the intervention (PTCA or CABG) conditioning on demographics, comorbidities, and outcomes using the *DE-SynPUF* data in comparison with actual Medicare 5% data.

When using the techniques described in "Synthetic File Creation Process" (see Section 6 on page 14) to create the synthetic files, the correlations among variables were altered. The synthetic processes were designed to try to preserve the relationships between demographic variables and clinical variables (diagnosis and procedures) in broad terms, such as gender to heart procedures. They were also designed to try to preserve the correlations of clinical variables, such as heart procedures to diabetes or high cholesterol. However, there were many processes used in creating the files, each designed to improve the confidentiality protection of the files, and by their nature, distort and alter correlations in the files. This resulted in *DE-SynPUF* where correlations among variables cannot be relied upon to be true.

The results in Table 11 and 12 show mixed results regarding the preservation of the correlation of demographics to specific clinical variables – PTCA and CABG. The gender distribution of synthetic beneficiaries with PTCA and CABG (shown in table 11) is very different than those of the actual beneficiaries. However, the odds ratios of gender (shown in table 12) are similar between the synthetic and real beneficiaries, indicating that preservation of correlations was only sometimes successful.

---

[6] PTCA http://www.webmd.com/heart-disease/angioplasty-with-stent-placement-for-heart-attack-and-unstable-angina
[7] CABG http://www.webmd.com/heart-disease/coronary-artery-bypass-surgery-for-coronary-artery-disease

Tables 11 also shows that the preservation of correlations between clinical variables was inconsistent as well. The rates of high cholesterol for CABG and PTCA beneficiaries are much higher in the synthetic sample than in the actual sample, while the rates of previous ischemic heart disease for the synthetic beneficiaries with CABG is similar to that of the actual beneficiaries with CABG. This illustrates that for this particular analysis some of the rates are closer to reality while others are not. Any specific analysis users would conduct with the *DE-SynPUF* (and compare with an actual sample of Medicare beneficiaries and their claims) could produce very different results (either worse or better) from this example.

Table 10. Definition of Diseases or Intervention

Variables of interest	ICD-9 Diagnosis codes	ICD-9 Procedures codes	DRG	HCPCS
CAD	414.00-414.07, 414.8x, 414.9x, 410.00-410.92, 412.xx, 411.0x-411.89, 413.0x-413.9x, V45.81, V45.82			(92980-92982, 92984, 92995, 92996, 33140, 33510-33514, 33516-33519, 33521-33523, 33533-33536) AND (99201-99205, 99212-99215, 99241-99245, 99354-99355, 99385-99387, 99395-99397, 99401-99404)
CABG		36.1x, 36.2x	106, 107, 109, 547-550	33510-33514, 33516-33519, 33521-33523, 33533-33536, 35600, 33572, S2205-S2209
PTCA		00.66, 36.01, 36.02, 36.05, 36.06, 36.07, 36.09	516, 517, 526, 527, 555-558	33140, 92980-92982, 92984, 92995, 92996
Depression	300.4x, 301.12, 309.0x, 309.1x, 311.xx			
Angina Pectoris	413			
Heart failure	398.91, 402.01, 402.11, 402.91, 428.0x-428.9x, 404.01, 404.11, 404.91, 404.03, 404.13, 404.93			
Ischemic Heart Disease	DX 410.00, 410.01, 410.02, 410.10, 410.11, 410.12, 410.20, 410.21, 410.22, 410.30, 410.31, 410.32, 410.40, 410.41, 410.42, 410.50, 410.51, 410.52, 410.60, 410.61, 410.62, 410.70, 410.71, 410.72, 410.80, 410.81, 410.82, 410.90, 410.91, 410.92, 411.0, 411.1, 411.81, 411.89, 412, 413.0, 413.1, 413.9, 414.00, 414.01, 414.02, 414.03, 414.04, 414.05, 414.06, 414.07, 414.10, 414.11, 414.12, 414.19, 414.2, 414.3, 414.8, 414.9	00.66, 36.01, 36.02, 36.03, 36.04, 36.05, 36.06, 36.07, 36.09, 36.10, 36.11, 36.12, 36.13, 36.14, 36.15, 36.16, 36.17, 36.19, 36.2, 36.31, 36.32		33510, 33511, 33512, 33513, 33514, 33515, 33516, 33517, 33518, 33519, 33521, 33522, 33523, 33533, 33534, 33535, 33536, 33542, 33545, 33548, 92975, 92977, 92980, 92982, 92995, 33140, 33141
Cardiac arrest	427.5x			
Hypertension	401.1x, 401.9x, 401.0x			
Diabetes	249, 250, 357.2, 362.01, 362.02, 366.41		637-639	
High cholesterol	272.0x			

Table 11. Bivariate Analysis of Intervention (PTCA or CABG) and Demographics, Comorbidities, and Outcomes—A Comparison of *DE-SynPUF* Data with Actual Medicare 5% Beneficiary Sample Data

Variables	*DE-SynPUF* CABG N	*DE-SynPUF* CABG %	*DE-SynPUF* PTCA N	*DE-SynPUF* PTCA %	Actual 5% Sample data CABG N[1]	Actual 5% Sample data CABG %[2]	Actual 5% Sample data PTCA N[1]	Actual 5% Sample data PTCA %[2]
Total (n, %)	**6,598**	**100**	**13,872**	**100**	**5,000**	**100**	**15,000**	**100**
Sex (n, %)								
Male	2,894	43.9	6,009	43.3	4,000	66	8,000	55
Female	3,704	56.1	7,863	56.7	2,000	34	7,000	45
Race (n, %)								
White	5,613	85.1	11,825	85.2	5,000	89	13,000	87
Black	652	9.9	1,299	9.4	*	6	1,000	8
Age group (n, %)								
younger than 70	2,405	36.5	4,884	35.2	2,000	37	5,000	36
70–79	2,272	34.4	4,960	35.8	2,000	45	6,000	38
80 and older	1,921	29.1	4,028	29.0	1,000	18	4,000	27
Comorbidity (n, %)								
High cholesterol	3,637	55.1	7,836	56.5	2,000	38	5,000	35
Hypertension	6,140	93.1	12,994	93.7	4,000	82	12,000	80
Previous Ischemic Heart Disease	4,855	73.6	10,498	75.7	4,000	75	10,000	64
Diabetes	5,499	83.3	11,703	84.4	2,000	44	6,000	43
Outcomes (n, %)								
Angina Pectoris	1,195	18.1	2,517	18.1	1,000	12	3,000	20
Heart Failure	3,505	53.1	7,536	54.3	2,000	41	5,000	36
Ischemic Heart Disease	4,723	71.6	10,228	73.7	5,000	90	12,000	83
Angina Pectoris, Heart Failure, Ischemic Heart Disease, or Cardiac Arrest	5,287	80.1	11,370	82.0	5,000	91	13,000	85

NOTE: [1] Rounded to the nearest multiple of 1,000.
[2] Rounded to the nearest whole numbers. Percentages may not add up to 100% due to editing to prevent disclosure.

Table 12. Heart Failure as an Outcome of the Intervention (PTCA or CABG) Conditioning on Demographics and Comorbidities—A Comparison of the *DE-SynPUF* Data with Actual Medicare 5% Beneficiary Sample Data

	DE-SynPUF OR	DE-SynPUF SE	Actual 5% Sample OR[1]	Actual 5% Sample SE
Outcome: Any Heart Failure				
Procedure (Reference group = Angioplasty)				
Bypass	0.98	0.03	1.21	0.04
Sex (Reference group = Male)				
Female	1.04	0.03	1.09	0.03
Race (Reference group = White)				
Black	1.06	0.05	1.38	0.08
Other	0.98	0.06	1.00	0.08
Age group (Reference group = younger than 70)				
70–79	1.04	0.04	1.15	0.04
80 or older	1.26	0.05	1.84	0.08
High cholesterol (Reference group = No)				
Yes	1.20	0.04	0.92	0.03
Hypertension (Reference group = No)				
Yes	1.47	0.10	1.44	0.06
Ischemic Heart Disease (Reference group = No)				
Yes	2.17	0.08	1.65	0.06
Diabetes (Reference group = No)				
Yes	2.06	0.09	1.54	0.05
Number of Cases	**20470**		**20,000**[2]	

NOTE: OR: Odds Ratio
SE: Standard Error
[1] Noise has been added to the regression coefficients for the actual 5% sample to ensure confidentiality of the beneficiaries. All coefficients are within two standard errors of the actual point estimates.
[2] Rounded to the nearest 100.

II. An HSR Example

In this basic HSR example, we were interested in predicting total inpatient expenditures in 2008 among beneficiaries who had at least one inpatient claim in 2008. Demographics (sex, age, race/ethnicity) and chronic conditions (Alzheimer or related disorders or senile, heart failure, chronic kidney disease, cancer-breast/colorectal/prostate/lung, chronic obstructive pulmonary disease [COPD], depression, diabetes, ischemic heart disease, osteoporosis, rheumatoid arthritis or osteoarthritis [RA/OA], and stroke/transient ischemic attack) were the covariates. We modeled the expenditures using a generalized linear model (GLM) with the gamma family and log link. From the *2008–2010 DE-SynPUF*, we used inpatient annual Medicare reimbursement amount, inpatient annual beneficiary responsibility amount, inpatient annual primary payer reimbursement amount, demographic variables, and chronic condition indicators. We merged the *DE-SynPUF*'s inpatient and beneficiary data to identify those who had at least one inpatient claim. Here, we present three tables with results from analyses using the *DE-SynPUF* and the same analyses computed on an actual Medicare 5% beneficiary sample.

Table 13 shows the univariate analysis of total inpatient expenditure, demographics, and chronic conditions using the *DE-SynPUF* data in comparison with actual Medicare 5% data. Table 14 shows the modeling results of total inpatient expenditure on demographics and chronic conditions comparing the *DE-SynPUF* data with actual Medicare 5% data. Table 15 shows the predicted mean total inpatient expenditure (margins) for covariates comparing the *DE-SynPUF* data with actual Medicare 5% data.

When using the techniques described in "Synthetic File Creation Process" (see Section 6 on page 14) to create the synthetic files, the correlations among variables were altered. For example, in Table 13, although the distributions of sex seem similar in the *DE-SynPUF* and the actual 5% sample, the prevalence rates of chronic conditions show much difference in the *DE-SynPUF* and the actual 5% sample. This illustrates that for this particular analysis some of the rates are closer to reality while others are not. Any specific analysis users would conduct with the *DE-SynPUF* (and compare with an actual sample of Medicare beneficiaries and their claims) could produce very different results (either worse or better) from this example.

Table 13. Univariate Analysis of Total Inpatient Expenditures, Demographics, and Chronic Conditions—A Comparison of the *DE-SynPUF* Data with Actual Medicare 5% Beneficiary Sample Data

Variables	*DE-SynPUF* N or Mean	*DE-SynPUF* % or SD	Actual data N[1] or Mean[1]	Actual data %[2] or SD[1]
Total (n, %)	**315,495**	**100**	**372,000**	**100**
Sex (n, %)				
Male	135,899	43.1	162,000	43
Female	179,596	56.9	210,000	57
Race (n, %)				
White	265,962	84.3	313,000	84
Black	33,785	10.7	41,000	11
Other	15,748	5	18,000	5
Chronic conditions (n, %)				
Alzheimer or related disorders or senile	163,582	51.9	81,000	22
Heart Failure	237,351	75.2	145,000	39
Chronic Kidney Disease	185,728	58.9	118,000	32
Cancer-Female breast/colorectal/prostate/lung	60,260	19.1	38,000	10
Chronic Obstructive Pulmonary Disease (COPD)	161,659	51.2	102,000	27
Depression	157,474	49.9	96,000	26
Diabetes	250,193	79.3	142,000	38
Ischemic Heart Disease	274,955	87.2	205,000	55
Osteoporosis	104,951	33.3	59,000	16
Rheumatoid Arthritis or Osteoarthritis (RA/OA)	113,983	36.1	115,000	31
Stroke/Transient Ischemic Attack	62,518	19.8	45,000	12
Total inpatient expenditures in 2008 (mean, SD)	**$18,766**	**$19,344**	**$18,000**	**$26,000**

NOTE: SD: Standard Deviation
[1] Rounded to the nearest 1,000.
[2] Rounded to the nearest whole numbers. Percentages may not add up to 100% due to editing to prevent disclosure.

Table 14. Modeling Results of Total Inpatient Expenditure on Demographics and Chronic Conditions—A Comparison of the *DE-SynPUF* Data with Actual Medicare 5% Beneficiary Sample Data

	DE-SynPUF Coef.	*DE-SynPUF* SE	Actual data Coef.[1]	Actual data SE
Total inpatient expenditure in 2008				
Sex (Reference group = Male)				
Female	-0.03	0.00	-0.14	0.00
Race (Reference group = White)				
Black	0.02	0.01	0.16	0.01
Other	0.00	0.01	0.19	0.01
Chronic conditions				
Alzheimer or related disorders or senile (Reference group = No)				
Yes	0.04	0.00	-0.08	0.01
Heart Failure (Reference group = No)				
Yes	0.16	0.00	0.24	0.01
Chronic Kidney Disease (Reference group = No)				
Yes	0.41	0.00	0.40	0.01
Cancer-Female breast/colorectal/prostate/lung (Reference group = No)				
Yes	0.18	0.00	0.13	0.01
Chronic Obstructive Pulmonary Disease (COPD) (Reference group = No)				
Yes	0.17	0.00	0.16	0.01
Depression (Reference group = No)				
Yes	0.02	0.00	0.22	0.01
Diabetes (Reference group = No)				
Yes	0.06	0.00	0.09	0.00
Ischemic Heart Disease (Reference group = No)				
Yes	0.19	0.01	0.11	0.00
Osteoporosis (Reference group = No)				
Yes	0.01	0.00	0.04	0.01
Rheumatoid Arthritis or Osteoarthritis (RA/OA) (Reference group = No)				
Yes	0.09	0.00	0.05	0.00
Stroke/transient Ischemic Attack (Reference group = No)				
Yes	0.16	0.00	0.23	0.01
Number of Cases	**315,495**		**372,000[2]**	

NOTE: SE: Standard Error

[1] Noise has been added to the regression coefficients for the actual 5% sample to ensure confidentiality of the beneficiaries. All coefficients are within two standard errors of the actual point estimates.

[2] Rounded to the nearest 1,000.