Data Entrepreneurs' Synthetic PUF: A Working PUF as an Alternative to Traditional Synthetic and Non-synthetic PUFs

J.M. Borton, A. T-C. Yu, A.M. Crego, A.C. Singh, M.E. Davern, and E. Hair
NORC at the University of Chicago, Chicago, IL 60603
borton-joshua@norc.org;

Abstract

The nature of Medicare Claims data makes it infeasible to apply usual methods of creating synthetic or non-synthetic PUFs due to data complexity, numerous identifying variables (IVs), and the difficulty of computing disclosure risk under any assumed intruder IV knowledge as knowledge may increase over time due to growing public availability of personal information. In view of this, we consider a two-prong strategy: creating a working PUF with high confidentiality at the cost of analytic utility, coupled with DUA-controlled access to microdata for testing the applicability of procedures developed for the working PUF in final analysis. The working PUF -- termed as data entrepreneurs' synthetic PUF (DE-SynPUF) -- has high pseudo-analytic utility in that it retains the original database structure and is thus useful to data entrepreneurs for application development and for researchers in training and some familiarization with the data on which the PUF is based. The DE-SynPUF was created by treating beneficiaries and individual claims, with no explicit preservation of intra-claim relationships. Moreover, all claims were subject to *post hoc* treatment in order to reduce risk when k-anonymization is used for de-identification. An application of DE-SynPUF to 2008-10 beneficiaries and claims data is presented.

Key Words: Synthetic and Non-synthetic PUFs; Dynamic Intruder Knowledge of Identifying Variables; Working PUF with High Confidentiality and Pseudo Analytic Utility

1. Infeasibility of Traditional Synthetic and Non-synthetic Methods

Usual methods of creating synthetic or non-synthetic PUFs for high confidentiality and analytic utility of data may not be appropriate or effective for Medicare Claims data. The complexity of the data in terms number and variety of variables, along with many relationships between all these variables across time make the application of traditional methods of disclosure protection very difficult, especially if one wants to preserve any kind of analytic utility. The main reason is the existence of a large number of related variables even for a single year let alone across multiple years from linked claims at the beneficiary level. Synthetic methods require disclosure-treatment of values of indirect identifying variables (IVs) for all records via parametric or nonparametric modeling of joint distributions for preserving relationships. Such modeling is difficult when dealing with a large number of IVs comprising discrete and continuous ones; see Raghunathan, Reiter and Rubin (2003), Reiter (2005), and Drechsler and Reiter (2010) among others.

Non-synthetic methods are also not quite suitable because of the frequent possibility of population uniques from too many unique combinations of IVs in the sample even if the intruder has knowledge of only a modest set of IVs about a target but longitudinally. The resulting preponderance of records at risk would imply treating almost all the beneficiary

records as in synthetic PUFs. However, by definition, for non-synthetic PUFs in the interest of analytic utility, only a small fraction of records (typically those with unusual profiles) are treated via perturbation or suppression of IVs or suppression of the whole record. Moreover, there is the problem of choosing a suitable set of IVs assumed to be known to the intruder because of its dynamic nature. That is, the disclosure risk believed to be small at the time of PUF creation may no longer be small over time due to increased intruder knowledge; see Singh (2009) and Borton, et al. (2011)

There is also the problem of the perception of disclosure, even when due to false claims of disclosure. With any unit level PUF where unit denotes a beneficiary in the case of claims data, it is difficult to reject with full confidence an intruder's assertion about even a false perception of identity disclosure of a target when the treated record might appear to match target's values of IVs. This problem is of concern with highly sensitive data such as Medicare claims, and can arise when methods designed to preserve utility are used and result in disclosure safe data that may appear to be disclosive to an intruder.

2. Two Prong strategy for Creating Unit Level PUFs

In view of the above mentioned problems about preserving both confidentiality and analytic utility in creation of usual synthetic and non-synthetic PUFs for Medicare Claims data at the beneficiary level, it is clear that a new strategy is needed. We consider a two-prong strategy with one prong involving the creation of a working PUF with very high confidentiality at the cost of true analytic utility using an aggressive nonrandom (or *ad hoc* random) disclosure treatment. This file has high pseudo analytic utility in that it reflects the true complexity, structure, and variability in the underlying data both at the beneficiary and claim level. The structure of the data has been very accurately preserved at the cost of interpretability so that it might be useful for data entrepreneurs for software and application development and training purposes for researchers. The data functions the same as the real data, but the results of any analysis are not reflective of the Medicare Population. To compensate for the loss of true analytic utility in the working PUF, the other prong in the two-prong strategy supplements by means of DUA-controlled access to microdata, under the assumption that programs and procedures developed for the working PUF remain applicable. A preferred cost and time saving alternative would be to use a disclosure-safe analysis tool not requiring a DUA but only a query-based access to the microdata through remote analysis servers. A possible alternative in this regard is discussed in Singh, Borton, and Crego (2012) and Singh, Borton, Davern, and Lin (2013).

3. Data Entrepreneurs' Synthetic PUF

With the revised goal of creating a working PUF and with a random sample of 2008 Medicare beneficiaries whose claims were tracked over 2008-2010, a new method termed "Data Entrepreneurs' Synthetic PUF" (DE-SynPUF) was developed in which beneficiaries and all claims records were treated in a piecewise manner taking one claim at a time with no built-in mechanism for preserving joint relationships between variables in order to reduce disclosure risk. Moreover, to avoid problems of perceived disclosure for unit level PUFs, all records were subject to *post hoc* treatment using perturbation or suppression and checked for disclosure by using k-anonymization type rules (FCSM, 2005) on subsets of IVs where all variables were regarded as potential IVs. Data from multiple donors for variables within and across claims were used to create synthetic beneficiaries so that combinations of variables observed in the DE-SynPUF could not lead to the disclosure that a particular beneficiary even contributed to the DE-SynPUF,

much less could an attribute of a particular beneficiary be disclosed. Further, beneficiary-level characteristics not having to do with claims, that is, geographic and demographic information, were treated in a similar fashion, 'constructing' synthetic beneficiaries out of multiple other beneficiaries. Further still, coarsening of important variables was employed. This led to a very conservative stance on disclosure risk tolerance and a highly safe PUF because subsets of variables created at the unit or synthetic beneficiary level were essentially at an aggregate level because they had common values over a corresponding group of donor beneficiaries. When needed, variables were syntheticly created by drawing from observed (and k-anonymized) univariate probability distributions. Thus commonly used methods for both synthetic and non-synthetic methods were used in combination. This resulted in a database whose structure is very similar to the CMS limited data sets, though with a smaller number of variables, some of which are coarsened or otherwise treated.

Synthetic beneficiaries and their claims are based on real 'seed' beneficiaries from a 5% sample of beneficiaries enrolled in Medicare in 2008. Tailored hotdecking procedures were used to find donors for beneficiary-level variables and individual claims. These methods create a 'synthetic' set of beneficiary characteristics and longitudinal; i.e., across claim profiles. Additional methods are then used to change variables within a single claim, ensuring that the variables associated with a single synthetic claim do not match those of a rare claim in the sample or the population.

While hot decking altered much information among beneficiary-level and claims records (including all four claim types), strategies were employed to attempt to alter high-level information as little as possible. For example, here are before-and-after comparisons of gender, race, and year-of-birth:

Table 1: Gender Comparison

GENDER	DE-SynPUF (%)	2008 5% Sample (%)
Male	44	45
Female	56	55

Table 2: Race Comparison

RACE	DE-SynPUF (%)	2008 5% Sample (%)
White	83	83
Black	11	10
Other	4	4
Hispanic	2	2

Table 3: Year of Birth Comparison

Year-of-Birth	DE-SynPUF (%)	2008 5% Sample (%)
post-1973	5	5
1964-1973	8	8
1954-1963	13	12
1944-1953	16	15
1939-1943	19	19
1934-1938	24	24
1929-1933	7	7
1924-1928	5	5
1919-1923	2	3
pre-1919	1	1

Also, the pattern for claim-making by percentage of overall beneficiaries was only marginally altered even though a substantial amount of information was altered behind the scenes both in terms beneficiary profiles and beneficiaries' claims:

Table 4: Claim Type Distribution Comparison

	DE-SynPUF (%)	2008 (%)	DE-SynPUF (%)	2009 (%)	DE-SynPUF (%)	2010 (%)
Inpatient	14	16	16	15	11	15
Outpatient	51	50	63	50	49	50
Carriers	73	70	80	70	76	70
PDE	63	53	79	56	74	57

Note: Percent of beneficiaries with at least one claim of specific claim types.

Also, mean reimbursements for each claim type was not altered much overall even while much information could change at the level of the beneficiary. So while some differences are starting to appear chiefly due top-coding of extreme values when making the DE-SynPUF, differences are not such that the data cannot be used for training and developing, and then applied to the DUA-controlled database.

Beneficiary-level indicators of a limited number of chronic conditions were created in the DE-SynPUF. Differences in their generation over and against what is actually done in the Chronic Conditions Warehouse on non-synthetic Medicare data are that, in the DE-SynPUF, each year is treated individually and no chronic condition is developed looking back over previous years, and in the DE-SynPUF, a beneficiary had to have the diagnosis trigger condition(s) hold within each year for respective yearly indicators.

For a complete list of variables and tables included in the DE-SynPUF please see the codebook which can be viewed at:http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/SynPUFs/Downloads/SynPUF_Codebook.pdf

Table 5: Mean Expenditure Comparison

	DE-SynPUF Mean	2008 Mean	DE-SynPUF Mean	2009 Mean	DE-SynPUF Mean	2010 Mean
Total Inpatient	2,550	2,850	2,500	3,050	1,450	3,050
Total Outpatient	850	1,150	1,050	1,250	600	1,300
Total Carrier	1,550	2,100	1,750	2,250	1,100	2,350
Total PDE	1,950	3,150	1,750	3,300	1,200	3,350

4. Empirical Results

The above method is applied to create a DE-SynPUF from Medicare claims data of a 5% sample of Medicare beneficiaries linked to their claims over three years, 2008-2010. The DE-SynPUF database is comprised of a Beneficiary Annual Summary File and four claim level files: Inpatient, Outpatient, Carrier, and Prescription Drug Events.

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 order 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.

4.1 A CER example (Comparative Effectiveness Research)

For this basic CER example, we were interested in the outcomes of angioplasty (*aka* percutaneous coronary interventions, denoted PTCA[1] hereafter) versus coronary artery bypass grafting (denoted CABG[2] 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 6 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 7 shows the results of bivariate analysis of intervention (PTCA or CABG) and demographics, comorbidities, and outcomes. Table 8 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.

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

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 Tables 2 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 7) is very different than those of the actual beneficiaries. However, the odds ratios of gender (shown in Table 8) are similar between the synthetic and real beneficiaries, indicating that preservation of correlations was only sometimes successful.

Table 7 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 6: 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

Variables of interest	ICD-9 Diagnosis codes	ICD-9 Procedures codes	DRG	HCPCS
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 7.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
Diabetes	5,499	83.3	11,703	84.4	2,000	44	6,000	43
Outcomes (n, %)								
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

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 8. 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.

4.2 An HSR Example (Health Service Research)

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 9 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 10 shows the modeling results of total inpatient expenditure on demographics and chronic conditions comparing the *DE-SynPUF* data with actual Medicare 5% data.

When using the techniques described earlier in this paper to create the DE-SynPUF, the correlations among variables were altered. For example, in Table 9, 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 9.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 10. 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				
Yes	0.04	0.00	-0.08	0.01
Heart Failure				
Yes	0.16	0.00	0.24	0.01
Chronic Kidney Disease				
Yes	0.41	0.00	0.40	0.01
Cancer-Female				
Yes	0.18	0.00	0.13	0.01
Chronic Obstructive Pulmonary				
Yes	0.17	0.00	0.16	0.01
Depression				
Yes	0.02	0.00	0.22	0.01
Diabetes				
Yes	0.06	0.00	0.09	0.00
Ischemic Heart Disease				
Yes	0.19	0.01	0.11	0.00
Osteoporosis				
Yes	0.01	0.00	0.04	0.01
Rheumatoid Arthritis or Osteoarthritis				
Yes	0.09	0.00	0.05	0.00
Stroke/transient Ischemic Attack				
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.

5. Concluding Remarks

The DE-SynPUF was created using a combination of synthetic, non-synthetic , and ad-hoc disclosure limitation methodologies. The goal of the DE-SynPUF process was to create a file with very high data confidentiality that retained the data structure of the source data. The preservation of the structure was at the expense of the analytic utility of the file. This makes the DE-SynPUF very different from most PUFs where often the analytic utility is preserved as much as possible, often resulting in a data structure that is very different from the source data it represents.

The resulting file gives the user an experience that is very similar to having access to the raw, claim level files. This allows the user to write code and develop products that can be used on the real data once access has be obtained via a DUA or some other type of remote data access. The primary (and intended) limitation of the DE-SynPUF is that while analysis can be executed, the results are not representative of the Medicare population, and one cannot make assertions or interpretations about the Medicare population based on the results of the analysis.

Acknowledgments and Disclaimer

The research in this article was supported in part by the Centers for Medicare and Medicaid Services under contract number 500-2006-00007I/#T0004 for the Medicare Claims CER Public Use Data Pilot Project. The views expressed in this article are those of the authors and do not necessarily reflect the views of the U.S. Department of Health and Human Services or the Centers for Medicare and Medicaid Services. The authors would like to thank Chris Haffer of CMS for his support and encouragement

References

Borton, J.M., Yu, A.T.-C., Crego, A.M., and Singh, A.C. (2011). Evaluation and Limitations of Disclosure-Treated Health Data Using Random Substitution and Sub-sampling. *Proceedings of Survey Research Methods Section*, American Statistical Association.

Drechsler, J. and Reiter, J.P. (2010) Sampling with Synthesis: A New Approach for Releasing Public Use Census Microdata, *JASA*, 105, 1347-1357.

Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of official Statistics*, 18, 531-544.

Reiter, J.P. (2005). Using CART to generate partially synthetic public use files. *Journal of Official Statistics*, 21,

Federal Committee on Statistical Methodology (2005). Report on Statistical Disclosure Limitation Methodology, Revised Working Paper No. 22 (1994), Bureau of the Census, US.

Singh, A.C. (2009). Maintaining analytic utility while protecting confidentiality of survey and nonsurvey data. *Journal of Privacy and Confidentiality* , Vol. 1, Number 2, 155-182.

Singh, A.C. and Borton, J.M. (2012). Aggregate Level PUF as a new alternative to the traditional unit level PUF for improving analytic utility and data confidentiality. *Proceedings of Survey Research Methods Section*, American Statistical Association.

Singh, A.C., Borton, J.M., and Crego, A. M. (2012). Generalized domain size for analysis restrictions with remote analysis servers. FCSM Proceedings, US census Bureau (www.fcsm.gov)

Singh, A.C., Borton, J.M., Davern, M., and Lin, Y. (2013). Query-Based PUF for Disclosure-Safe Remote Analysis from Medicare Claims Micro Data. *Proceedings of Survey Research Methods Section*, American Statistical Association.