Aggregate Level PUF with High Data Confidentiality and Utility for General Purpose Analyses from Medicare Claims Data

A.C. Singh[1], J.M Borton[1], E. Erdem[2], and Y. Lin[1]
[1]NORC at the University of Chicago, Chicago, IL 60603
[2]KPMG LLP, McLean, VA 22102
singh-avi@norc.org; borton-josh@norc.org; erkanerdem@kpmg.com; lin-yongheng@norc.org

Abstract

Creating a unit level PUF that is analytically useful and disclosure-safe is difficult due to the proliferation of publicly available indirect identifiers from various known and unknown sources that might exist at present or in future. In creating an aggregate level (AL-) PUF for Medicare Claims data, the data structure is transformed from a beneficiary level file with rows representing beneficiaries and columns analytic variables to a file with rows representing small clusters of beneficiaries called micro groups (MGs) and columns representing MG size and various domain means at the MG level (termed micro means--MMs) where domains are subpopulations of beneficiaries defined by variables corresponding to analytic goals. This allows information to be presented without sharing unit level data. Uncertainty in the MG size and associated MMs is introduced by random subsampling followed by weight calibration. The MGMM structure of AL-PUF is somewhat similar to the method of micro-aggregation for unit level PUFs where values of continuous variables deemed to be identifying are blurred by averaging over small clusters of observations based on similarity indices. However, the main difference is that in AL-PUF, MMs are provided for various domains defined by one or more variables and so joint relationships between variables are not distorted unlike the case of micro-aggregation. Moreover, as a result of subsampling, AL-PUF does not require the framework of identifying and sensitive variables used in traditional methods for creating unit level PUFs. For analysis domains, descriptive and analytic parameters of interest can be estimated using the weighted sums of products of suitable domain MMs and MG size over all MGs. Estimation of variance and covariance of point estimates can be obtained using essentially standard sampling methods because sampling errors introduced in MMs and MG size are due to multi-phase sampling. AL-PUF achieves high confidentiality by using MG sizes sufficiently small to reflect adequate uncertainty due to sampling errors but large enough to avoid problems with unit level PUFs. It has high analytic utility because analytic domains could be defined for any subset of variables of interest and the corresponding estimates remain approximately unbiased because uncertainty is only due to random subsampling. Examples from a 15% random sample of the 2010 Medicare claims data on chronic conditions are presented for comparisons between AL-PUF and an existing unit level PUF (termed chronic conditions PUF) based on k-anonymization and micro-aggregation.

Key Words: Unit Level vs. Aggregate Level; Micro Groups and Micro Means; Micro Aggregation;Subsampling

1. Introduction

Traditional methods of data input de-identification at the unit level for creating unit level PUFs are based on the identifying variable/sensitive variable (IV/SV) framework where IVs denote indirect (or quasi) identifiers (various subsets of which are assumed to be known to potential intruders) and SVs denote variables of interest to the intruders; see Duncan et al. (2011). In any de-identification approach, IVs are disclosure-treated either by nonsynthetic methods of perturbation and/or suppression or synthetic methods of partial synthesis via modeling. In view of the possibility that intruders might know unspecified subsets of more IVs than what was assumed, traditional methods of PUF creation using the IV/SV framework may no longer be safe. In other words, for a dataset with many analytic variables (or the dataset created by linking various databases cross-sectionally and longitudinally to obtain a rich analytic file), it is difficult to draw a line between IVs and SVs because the intruder knowledge could grow over time. It follows that to be conservative it is preferable not to assume the intruder knowledge to be static. It is remarked that the main reason for the inadequate disclosure-safety of traditional PUFs is that the joint information about too many variables associated with an individual or unit is released simultaneously after disclosure- treatment of IVs although users need only a subset of variables at a time for analysis.

The main purpose of this paper is to present an application of an aggregate-level PUF (AL-PUF) introduced by Singh and Borton (2012) as an alternative to traditional unit-level PUFs which can overcome concerns mentioned above to a great extent as it does not use the IV/SV framework. The example used for illustrating the application is based on a 15% sample of the 2010 CMS Medicare Claims data Basic Annual Summary File with information about chronic conditions, demography, reimbursement, and other variables as well as. An existing chronic conditions PUF (CC-PUF) constructed recently (Erkan et al., 2012; see also www.cms.gov under Research, Statistics, Data, and Systems). We present various analysis results showing comparison of estimates from CC-PUF and AL-PUF relative to results from the 15% sample regarded as true values.

2. The Example of Chronic Conditions PUF Creation as an Application to Medicare Claims Data

The existing *CMS CC- PUFs* represent 100% of the Medicare beneficiaries provided in the 100% Beneficiary Summary File for the reference year. The 100% Beneficiary Summary File is created annually and contains demographic, entitlement and enrollment data for beneficiaries who were:

- Documented as being alive for some part of the reference year of the Beneficiary Summary File, and
- Entitled to Medicare benefits during the reference year, and
- Enrolled in Medicare Part A and/or Part B for at least one month in the reference year.

Beneficiaries with 12 months of enrollment in Fee-for-Service (FFS) plans of Part A or Part B are separated from beneficiaries with less than 12 months of enrollment. Beneficiaries with less than 12 months of enrollment include:

- Beneficiaries who turned 65 in the calendar year,

- Beneficiaries who died during the calendar year, and
- Beneficiaries who switched in and out of Medicare Part C, or Medicare Advantage (MA) plans, during the calendar year.

Note also that the *CC- PUFs* include information from beneficiaries who are enrolled in Medicare on the basis of disability and End-Stage Renal Disease (ESRD).

The goal of CC-PUFs is to provide information about outcome variables such as types of health care utilization and reimbursements for various analysis domains defined by auxiliary variables given by age, gender, chronic condition (11 of them) , dual eligibility status (Medicaid and Medicare), and length of enrolment. Taking a conservative stance, all variables (auxiliary and outcome) are treated as potential IVs and even knowledge of the presence of a target in the database is deemed as disclosure. The de-identification of *CC- PUFs* uses k-anonymization (i.e., global recoding also known as generalization and local suppression) on categorical IVs and micro-aggregation on continuous IVs (i.e., report means over small groups of beneficiaries). In creating CC-PUFs, it turns out that after k-anonymization, all the beneficiaries could be classified into cells or profiles defined by the auxiliary variables, and therefore, the unit level PUF is essentially transformed into a cell or profile level data or a macro data with counts and magnitudes in the form of cell averages. Thus, each cell can be viewed as a micro-aggregate for reporting averages of continuous outcome variables such as utilization and reimbursement. Rules for minimum threshold for k-anonymization and micro-aggregation are described below.

First, some profiles are coarsened so that every profile contains at least 30 beneficiaries (those enrolled in Medicare Part A or Part B for at least one month in the calendar year). This was done by local suppression, that is, by making the actual value of some (6 out of 11) chronic conditions missing/blank. Hence, even though the entire list of chronic conditions is available in the file, the values for some of the chronic condition indicators are not provided for some of the profiles. Because of this step, six chronic conditions are suppressed for less than 0.5 percent of the beneficiaries in the *PUFs.*

Second, all cost and/or utilization measures are replaced with missing/blank if the number of beneficiaries for a particular block is less than 30. For example, if there are only 8 beneficiaries with enrollment in Part A for 12 months for a particular profile, then none of the cost and/or utilization measures are reported for that block. Note that the measures are available for other blocks (e.g., beneficiaries with enrollment in Part A for less than 12 months) in the same profile as long as they contain at least 30 beneficiaries.

Third, a cost and/or utilization measure is replaced with missing/blank if the number of beneficiaries with at least one claim (for the relevant service) is less than 11. For example, if there are only 8 beneficiaries with at least one inpatient admission in a particular block, then the cost and/or utilization measures for inpatient services are not reported. Note that the measures are available for other blocks in the same profile as long as there are at least 11 beneficiaries with a claim. This also applies to the variables that contain the average total Medicare reimbursement for each block (e.g., average Medicare payment per beneficiary for all Part A services for beneficiaries with enrollment in Part A for 12 months). That is, the calculation of every measure in the PUF is based on at least 11 beneficiaries with a claim associated with that measure. The number of enrolled beneficiaries in each block is not suppressed.

The construction of CC-PUFs is appealing because of its simplicity and ease in implementation. However, there may be several concerns. First, nonrandom suppression of data under k-anonymization may introduce serious bias especially for small domains defined for example by demographics and chronic conditions, although for large domains, estimates may behave reasonably well. Second, due to nonrandom suppression or missing data, standard software for analysis to deal with missing data are not applicable and standard errors of estimates are not available for making inference, in particular. Third, again due to local suppression of certain variables, it is not possible to produce lower dimensional or marginal tables from the tabular form of CC-PUF even for marginal tables which may not require any disclosure treatment by themselves.

Although the original CC-PUFs were created from the 100% Beneficiary summary file for the reference year, we recreated CC-PUF for a simple random sample of 15% beneficiaries for the year 2010 for ease in comparison with AL-PUF. The 15% sample is regarded as the population (termed untreated database) against which results from CC-PUF and AL-PUF are compared as shown in Tables 1-3 and discussed in Section 5. For this purpose, the analytic questions at the beneficiary level are based on the diabetes chronic condition and corresponding counts and re-imbursement amounts for domains or subpopulations defined by age, gender, and the presence of comorbidities due to other chronic conditions.

3. Aggregate Level PUFs: A Description

Here we briefly describe the method of AL-PUF as introduced by Singh and Borton (2012). It uses the framework of Micro Groups (MGs) and Micro Means (MMs) defined as follows.

3.1 Transformation from Unit Level to Aggregate Level

We divide the original unit level data (denote by s_0 --a 15% subsample of the full population or universe *U*) into small clusters or groups (termed MGs) of size 10-20 using a broad cross-classification of age (4 categories), race (4), sex (2) and dual status (2); i.e., 96 combinations although only 64 are nonempty. The stratum size varied from 3400 to 1.2M. The total sample size is about 7.2M and the number of MGs is 478594; i.e., about 480K. Beneficiaries in each stratum were serpentine-sorted in order using the variables sex(2), age(21), dual status(2), race(5) followed by enrollment status (none, part, full year) in Part A, B, D and then C. This was done to make the MGs somewhat homogeneous with respect to claims and expenditures. MGs should not be too small to avoid problems of unit level PUFs and not too large to reflect sufficient uncertainty due to subsampling explained in the Sub-section 3.2. The actual MG size is randomly chosen between 10 and 20 for each MG.

In the AL-PUF data structure, for each MG, MG size and MMs are presented for several outcome variables (z) for each analytic domain D of interest. For example, the domains could be defined by auxiliary variables (demography and chronic conditions) used alone or in combinations, and outcome variables (z) may correspond to simply domain indicator or different types of health care utilization and reimbursements. Thus the MGMM structure of AL-PUF consists of MGs as rows instead of individual beneficiaries, and MMs for each domain D along with MG size as columns instead of values of individual variables—auxiliary or outcome. By way of notation, we will denote

the estimated MG size or the population count based on the sample s_0 by $\hat{N}_{g(s_0)}$, MM for the variable z for domain D by $\hat{M}_{zg(D,s_0)}$ which is obtained as the ratio of the estimated MG total $\hat{T}_{zg(D,s_0)}$ and the MG size $\hat{N}_{g(s_0)}$, and the MM or micro-proportion for the special case of z being the domain indicator as $\hat{P}_{g(D,s_0)}$; all estimates are sample weighted after calibration to selected control totals from the population *U* in the interest of balancing the sample. We thus have an equivalent representation for estimated total $\hat{T}_{z(s_0,D)}$ and estimated count $\hat{N}_{D(s_0)}$for the domain D from the unit level data for the sample s_0 in terms of aggregate level data as

$$\hat{T}_{z(D,s_0)} = \sum_{g=1}^{G} \hat{N}_{g(s_0)}\, \hat{M}_{zg(D,s_0)} \;, \hat{N}_{D(s_0)} = \sum_{g=1}^{G} \hat{N}_{g(s_0)}\, \hat{P}_{g(D,s_0)} \,, \qquad (1)$$

where *G* is the total number of MGs.

3.2 Disclosure-Safety: Aggregate Level Transformation and Nested Subsampling

Although aggregate level transformation goes a long way in protecting individual information, there may be disclosure of z-values for beneficiaries in rare domains if an intruder computes the MG total $\hat{T}_{zg(D,s_0)}$; i.e., the product $\hat{N}_{g(s_0)} \times \hat{M}_{zg(D,s_0)}$ for the MG with a nonzero value of $\hat{P}_{g(D,s_0)}$. This problem can be alleviated by introducing uncertainty in the MG size $\hat{N}_{g(s_0)}$ by using a different estimate based on a random subsample. Similarly, computing the ratio $\hat{M}_{zg(D,s_0)}/\hat{P}_{g(D,s_0)}$ could lead to disclosure of z-values of beneficiaries in rare domains because denominators of the two terms cancel each other. Now, introducing uncertainty in the micro proportion $\hat{P}_{g(D,s_0)}$ by subsampling can overcome this problem. Similarly, the micro mean $\hat{M}_{zg(D,s_0)}$ may disclose z-values in s_0 for a beneficiary in a rare domain which can again be overcome by subsampling to introduce uncertainty. Thus we need three subsamples of s_0: $s_1$to obtain $\hat{M}_{zg(D,s_1)}$, s_2 to obtain $\hat{P}_{g(D,s_2)}$, and s_3 to obtain $\hat{N}_{g(s_3)}$. The subsamples are drawn in a nested manner $s_3 \subset s_2 \subset s_1 \subset s_0$ to yield unbiased estimates as

$$\hat{T}_{z(D,s_1,\, s_3)} = \sum_{g=1}^{G} \hat{N}_{g(s_3)}\, \hat{M}_{zg(D,s_1)} \;, \hat{N}_{D(s_2,s_3)} = \sum_{g=1}^{G} \hat{N}_{g(s_3)}\, \hat{P}_{g(D,s_2)} \,, \qquad (2)$$

using the conditioning arguments as in three-phase sampling. With suitable sampling rates (see next subsection), three-phase sampling can introduce sufficient desired uncertainty through sampling errors without introducing bias in domain counts and total estimates.

It may be noted that the MGMM structure of AL-PUF is somewhat similar to the method of micro-aggregation (Domingo-Ferrer and Torra, 2001) used for creating traditional unit level PUFs with some exceptions. First, there is no need of the IV/SV framework in AL-PUF because it is an aggregate level file providing information via MMs for analysis domains and incorporates uncertainty in MG level estimates for each domain via subsampling. Second, unlike micro-aggregation, there is no distortion of joint relationships between variables because domain can be defined by values of single or multiple variables. It may also be noted that the problem of possible dominance by an individual observation in an MG level MM is diffused by uncertainty in the MG size and the presence of a target in the MG due to subsampling making it difficult to reconstruct the MG total with any reasonable precision.

Sampling weights for each subsample should be calibrated to control totals obtained from s_0 which itself is calibrated to external control totals available for the population *U*. Calibration is useful for a good representation of the population in the sample as well as it can introduce more uncertainty in the MG size $\widehat{N}_{g(s_3)}$, micro-proportion $\widehat{P}_{g(D,s_2)}$, and the micro-mean $\widehat{M}_{zg(D,s_1)}$ at the MG level although at the higher domain D level, it is expected to introduce more stability. Table 4 presents a list of calibration controls (213 of them) based on two-factor crossed categories of demography and one-factor marginal variables of chronic conditions. Although all subsample weights are calibrated, it follows from equation (2) that the final estimates do not satisfy any of the calibration controls because they use a non-standard three-phase sample estimation in that a hybrid of estimates based on all the three subsamples is employed for protecting confidentiality.

3.3 Descriptive and Analytic Inference with AL-PUF

We observe that any parameter of interest in the form of a domain total can be easily estimated using formula (2) under the MGMM structure by defining domain-specific variables (auxiliary and outcome) and computing corresponding MMs. The estimate is unbiased because of the nested subsampling in the standard theory of multi-phase sampling. Specifically, we have

$$E(\hat{T}_{z(D,s_1,s_3)}) = E_1E_2E_3\left(\sum_{g=1}^{G}\widehat{N}_{g(s_3)}\,\widehat{M}_{zg(D,s_1)}\right) = E_1E_2\left(\sum_{g=1}^{G}E_3(\widehat{N}_{g(s_3)})\,\widehat{M}_{zg(D,s_1)}\right)$$

$$= E_1E_2\left(\sum_{g=1}^{G}\widehat{N}_{g(s_2)}\;\widehat{M}_{zg(D,s_1)}\right) = E_1\left(\sum_{g=1}^{G}E_2(\widehat{N}_{g(s_2)})\,\widehat{M}_{zg(D,s_1)}\right)$$

$$= E_1\left(\sum_{g=1}^{G}\widehat{N}_{g(s_1)}\;\widehat{M}_{zg(D,s_1)}\right) = E_1\left(\sum_{g=1}^{G}\hat{T}_{zg(D,s_1)}\right) = \sum_{g=1}^{G}E_1(\hat{T}_{zg(D,s_1)})$$

$$= \sum_{g=1}^{G}\hat{T}_{zg(D,s_0)} = \hat{T}_{z(D,s_0)} \qquad (3)$$

as desired where $E_1E_2E_3$ denote expectation operators corresponding to the three stages of random subsampling conditional on the previous higher phase. Similarly, $\widehat{N}_{D(s_2,s_3)}$ is unbiased for $\widehat{N}_{D(s_0)}$. Observe that use of a lower level subsample to estimate the MG size and not the higher level MM in constructing estimators $\hat{T}_{z(D,s_1,s_3)}$ and $\widehat{N}_{D(s_2,s_3)}$ under the nested subsampling structure makes the above argument for unbiasedness go through. Strictly speaking we only get approximately unbiased estimates because the sampling weights in the above estimates are calibrated and render the estimates nonlinear. For nonlinear parameters such as means, ratios, proportions, and odds ratios, we can obtain only approximately unbiased estimates with AL-PUF as is the case with the original data.

For variance estimation of the domain total estimator $\hat{T}_{z(D)}$, we consider

$$Var\left(\hat{T}_{z(D,s_1,s_3)}\right) = V_{123}\left(\hat{T}_{z(D,s_1,s_3)}\right) = E_{123}(\hat{T}_{z(D,s_1,s_3)} - E_{123}\left(\hat{T}_{z(D,s_1,s_3)}\right))^2$$

$$= E_1V_{23}(\hat{T}_{z(D,s_1,s_3)}) + V_1E_{23}(\hat{T}_{z(D,s_1,s_3)})$$

$$= E_1E_2V_3\left(\hat{T}_{z(D,s_1,s_3)}\right) + E_1V_2E_3\left(\hat{T}_{z(D,s_1,s_3)}\right) + V_1E_2E_3\left(\hat{T}_{z(D,s_1,s_3)}\right)$$

$$= E_{12}V_3\left(\sum_{g=1}^{G}\hat{N}_{g(s_3)}\,\hat{M}_{zg(D,s_1)}\right) + E_1V_2\left(\sum_{g=1}^{G}\hat{N}_{g(s_2)}\,\hat{M}_{zg(D,s_1)}\right) + V_1\left(\sum_{g=1}^{G}\hat{N}_{g(s_1)}\,\hat{M}_{zg(D,s_1)}\right) \quad (4)$$

Now, standard design-based formulas can be used to obtain approximately unbiased and consistent estimates of the three variances of estimated totals at different phases in the above expression. However, to compute variance estimates with AL-PUF, we need to define suitable outcome variables such as squares and cross-products of variables under study and required domains to compute the necessary domain totals from the MGMM structure. For nonlinear estimators, Taylor linearization can be used to express the estimator approximately as a linear estimator and then the above formula is used. Thus for descriptive inference, in general, we can obtain point and variance estimates.

Similarly for analytic inference with linear models, where closed analytic forms of estimators are available, we can obtain point and variance estimates as in the case of descriptive inference. However, for nonlinear models, where closed form expressions of estimators are not available, use of MGMM structure to compute estimates iteratively is time consuming and may not be appealing in practice. This is a practical limitation of AL-PUF but it is useful for general purpose analyses with high utility and confidentiality (see next section for risk and utility measures). A query-based method such as Q-PUF (Singh et al., 2013) is useful for complex analytic problems.

For regression model diagnostics, a modified residual analysis can be performed by defining domains or intervals from the predicted values on the x-axis, and computing domain means of unit-level residual using the MGMM structure. Thus estimated domain means of residuals can be plotted against the actual domain means of predicted values as an alternative to the original unit level residual plot which may not be safe.

3.4 Utility Tool for Computing Estimates for Arbitrary User-specified Domains

An important advantage of AL-PUF over Q-PUF is that estimates for any user-specified domains can be obtained regardless of how rare the domains are. Disclosure protection against rare domains is built-in though sampling errors in that estimates for rare domains would not be of much analytic use due to high sampling errors. However, it is not practical for the data producer to anticipate a multitude of possible domains users might be interested in and provide in advance all the necessary MMs for each such domain. Nevertheless, it is possible to anticipate a set of domains commonly used by analysts and a file for each such domain can be part of a basic AL-PUF dataset prepared by the data producer. This can then be supplemented by a utility tool that can communicate with the original microdata through a user interface and produce in real time all the MMs required for a new user-specified domain using the simple MGMM structure already in place. This is not expected to be time consuming because the MM calculation requires only simple operations of addition and division for each MG.

4. Metrics for Disclosure Risk and information Loss

Let subsampling rates for s_1, s_2, and s_3 be denoted respectively f_1, f_2, and f_3. These rates are chosen such that both risk and information loss are below pre-specified acceptable levels. To measure disclosure risk, we define several events as follows. For a given small positive number τ used as a threshold for the absolute relative error, we define for each MG g,

$$\text{A}= \left[\left|\widehat{N}_{g(s_1)}/\widehat{N}_{g(s_0)} - 1\right| < \tau\right],\ \text{B} = \left[\left|\widehat{N}_{g(s_2)}/\widehat{N}_{g(s_1)} - 1\right| < \tau\right]$$
$$\text{C}= \left[\left|\widehat{N}_{g(s_3)}/\widehat{N}_{g(s_1)} - 1\right| < \tau\right],\ \text{D} = \left[\left|\widehat{N}_{g(s_2)}/\widehat{N}_{g(s_0)} - 1\right| < \tau\right]$$
$$\text{E}= \left[\left|\widehat{N}_{g(s_3)}/\widehat{N}_{g(s_2)} - 1\right| < \tau\right],\ \ \text{F} = \left[\left|\widehat{N}_{g(s_3)}/\widehat{N}_{g(s_0)} - 1\right| < \tau\right]$$

Then, quantiles of the distribution of Pr(A) over all MGs, g=1,…,G, give risk metrics (denoted by δ's) for the potential disclosure problem when $\widehat{M}_{zg(D,s_1)}$ is very close to $\widehat{M}_{zg(D,s_0)}$. Actually, this probability is only an upper bound because we need both $\widehat{N}_{g(s_1)}$ and $\widehat{T}_{g(D,s_1)}$ close to $\widehat{N}_{g(s_0)}$ and $\widehat{T}_{g(D,s_0)}$, but for simplicity to avoid computation of domain-specific risk metrics, we only look at MG sizes. The reason for introducing nested subsampling was that we want various probabilities of events AB, DE, ABE, and ABCDEF sufficiently small after pre-multiplication by the sampling fractions f_1 f_2 to account for the uncertainty that a beneficiary with a rare profile may not be in samples s_1 or s_2 in order to be at risk. In addition to being upper bounds as mentioned above, these risk measures are rather conservative because they also need to be pre-multiplied by f_0 to account for the uncertainty that the target may not even be in the original dataset s_0.

To estimate above risk measures, several simulated samples are drawn from s_0 using stratified simple random sampling without replacement, and for each sample, occurrences or non-occurrences of above events are recorded. For the 2010 CMS Medicare Data application, Table 5 shows the risk measures ($\delta's$) based on 500 simulations obtained for 50% and 75% quantiles when the sampling rates were set at (.9,.9,.6) and (.9,.9,.2) and the threshold τ for the absolute relative error was set at .05, .1, and .2. The simulations were restricted so that for each MG, $\widehat{N}_{g(s_2)}$ is not zero; i.e., each MG is populated by the sample s_2. The column under (.9,.9,.2) with τ=.10, and 75% quantile seems to provide a reasonable guide for choosing sampling rates. This choice was used for producing AL-PUF estimates reported in the next section.

The above choice of sampling rates was also found to be reasonable for controlling information loss. For measuring information loss, we considered 180 fairly small domains defined by a cross-classification of age (6 categories) by race (5) by gender (2) by diabetes (2) and by COPD (2). For each domain d, absolute relative errors of domain count and domain total expenditure estimates were computed to see if they are above the threshold τ or not for each simulation and the corresponding probability was computed. Specifically we consider the events for defining information loss as

$$\left[\left|\widehat{N}_{D(s_2,s_3)}/\widehat{N}_{d(s_0)} - 1\right| > \tau\right] \text{ and } \left[\left|\widehat{T}_{z(D,s_1,\,s_3)}/\widehat{T}_{zd(s_0)} - 1\right| > \tau\right]$$

Table 6 shows measures of information loss (denoted by ε's) defined by quantiles (50% and 75%) of these empirical probabilities over 180 domains. It follows that the choice of (.9,.9,.2) provides a reasonable balance between control of disclosure risk and information loss.

5. Comparison of Estimates between CC-PUF and AL-PUF for the 2010 Medicare Claims Data

In Tables 1-3, only descriptive comparisons of point estimates are presented. As expected, CC-PUF shows downward bias in total estimates for all the study variables considered. However, in estimating mean expenditure, the bias could be upward or downward. For small domains (Table 2) and even smaller domains (Table 3), there are more downward biases in total count and expenditure estimates.

In tables 3.1 through 3.6 it is interesting to note that the amount of error remains generally constant across different comorbidities for the CC-PUF method while error generally decreases with the domain size for AL-PUF values. This indicates that estimates for CC-PUF for smaller domains contain bias, while estimates for AL-PUF do not have bias, though they do have more estimation error than AL-PUF estimates for larger domains.

The acronyms for co-morbidities used in Table 3 are: IHD: Ischaemic Heart Disease, CHF: Congestive Heart Failure, and CAN: Cancer.

6. Concluding Remarks

In this paper an important application of the AL-PUF method to the 2010 CMS Medicare Claims data was presented and compared with a much simpler method of CC-PUF currently used. It was found as expected that for large domains, the two methods provide similar point estimates of domain counts (diabetes) and totals (reimbursement) although there is some downward bias with CC-PUF but for small domains, bias in CC-PUF can be serious. Also with CC-PUF, it is not possible with standard methods to account for bias due to nonrandom suppression in estimating variance required for making inference.

The aggregate level data structure of AL-PUF consists of MGs and domain MMs and nested subsampling is used to introduce uncertainty in domain MMs, domain micro-proportion, and the MG size for each MG. Subsampling rates are chosen such that both disclosure risk and information loss are controlled at suitable levels, thus providing high confidentiality and utility. The AL-PUF method is somewhat similar to the method of micro-aggregation but it is different in several important ways: first, it is not at the unit level because it provides information through domain means at the MG level; second, it does not use the framework of IV/SV because it introduces uncertainty through subsampling in domain means and the MG size; and third, it preserves joint relationships between variables because domains can be defined by one or more variables.

With AL-PUF, both descriptive and analytic inferences can be made. Suitable variance estimates can be obtained from standard multi-phase sampling methods. An approximate residual analysis for model diagnostics can also be performed by transforming unit level residuals to an aggregate level using the MGMM framework of AL-PUF when domains are defined by small bins on the x-axis obtained by partitioning the predicted values.

For arbitrary analytic variables once the desired domains are specified, corresponding MMs can be computed to estimate domain totals. This type of specification is needed for variance and covariance estimation for squares and cross-products of variables. However,

it may not be practical for the data producer to anticipate in advance various possible domains that might be of interest to users. This necessitates the provision of a utility tool to supplement a basic AL-PUF data file so that users can compute estimated totals for arbitrarily specified domains in real time. It follows that AL-PUF may be most suitable for general purpose analyses but not for complex ones involving nonlinear models. The method of Q-PUF, on the other hand, proposed for query-based systems is suitable for complex analyses but it does not allow for specification of arbitrary analysis domains although all commonly used analysis domains can be allowed.

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

Domingo-Ferrer, J., and Torra, V. (2001). A quantitative ecompariosn of disclosure control methods for microdata. In Zayatz, L., Doyle, P., Theeuwes, J., and Lane, J. (Eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland: Amsterdam, pp. 111-133.

Duncan, G.T., Elliott, M., and Salazar-González, J.-J. (2011). *Statistical Confidentiality,: Principles and Practice.* Springer: New York.

Erkan, E., Prada, S., and Haffer, S.C. (2013). Medicare payments: How much do chronic conditions matter?, Medicare and Medicaid Research Review, Vol 3, No. 2, 1-15. (http://www.cms.gov/mmrr/Downloads/MMRR2013_003_02_b02.pdf)

Singh, A.C., Borton, J.M., Davern, M. E., and Lin, Y. (2013). Query-based PUFs for Disclosure-Safe Remote Analysis from Medicare Claims Micro Data, *ASA Proc., Surv. Res. Meth. Sec.,*

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

Tables and Figures

Tables 1 through 3 are representative sub sets of all tables created to observe the differences between CC-PUF and AL-PUF treated data and the original untreated data. In the interest of space these tables are shown for a limited number of cohorts.

Table 1.1 Diabetes Counts and Reimbursement Amounts for 2010 Medicare Beneficiaries
(Domain: Nondual eligible, Part A)

	ALL			Diabetes=Yes			Diabetes=No		
	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb
Untreated	$10,000,000,000	5,000,000	$3,000	$5,000,000,000	1,000,000	$6,000	$8,000,000,000	4,000,000	$2,000
CC-PUF	-$1,521,784,820	-1,292,119	$441	-$635,381,000	-85,465	-$178	-$886,403,820	-1,206,654	$433
AL-PUF	-$12,855,313	-1,363	-$2	-$10,569,354	-633	-$7	-$2,285,959	-730	$0

Table 1.2 Diabetes Counts and Reimbursement Amounts for 2010 Medicare Beneficiaries
(Domain: Nondual eligible, Part B)

	ALL			Diabetes=Yes			Diabetes=No		
	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb
Untreated	$10,000,000,000	5,000,000	$3,000	$5,000,000,000	900,000	$5,000	$9,000,000,000	4,000,000	$2,000
CC-PUF	-$1,030,728,421	-1,276,848	$759	-$384,863,754	-84,919	$92	-$595,795,774	-1,191,929	$796
AL-PUF	-$10,853,470	-1,663	-$1	$496,252	-483	$3	-$11,349,722	-1,180	-$2

NOTE: For disclosure concerns, all entries in the tables have been rounded to only one significant digit for untreated data or the 15% sample. CC-PUF and AL-PUF results are shown as differences to illustrate the comparison without sharing actual estimates.

Table 2.1 Diabetes Counts and Reimbursement Amounts for 2010 Medicare Beneficiaries
(Domain: M under 65, Nondual eligible, Part A)

	ALL			Diabetes=Yes			Diabetes=No		
	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb
Untreated	$700,000,000	300,000	$2,000	$300,000,000	60,000	$6,000	$300,000,000	300,000	$1,000
CC-PUF	-$56,069,273	-54,701	$216	-$29,468,508	-5,387	$49	-$26,600,765	-49,314	$161
AL-PUF	-$9,995,581	635	-$36	-$4,298,841	91	-$87	-$5,696,740	544	-$24

Table 2.2 Diabetes Counts and Reimbursement Amounts for 2010 Medicare Beneficiaries
(Domain: M 65-69, Nondual eligible, Part A)

	ALL			Diabetes=Yes			Diabetes=No		
	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb
Untreated	$900,000,000	600,000	$1,000	$400,000,000	100,000	$4,000	$500,000,000	500,000	$1,000
CC-PUF	-$72,628,623	-138,887	$287	-$38,610,944	-10,295	$26	-$34,017,679	-128,592	$240
AL-PUF	$18,327,150	1,132	$28	$8,054,109	154	$71	$10,273,041	978	$19

Table 2.3 Diabetes Counts and Reimbursement Amounts for 2010 Medicare Beneficiaries
(Domain: M 70-74, Nondual eligible, Part A)

	ALL			Diabetes=Yes			Diabetes=No		
	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb	Tot Reimb	Count	Mean Reimb
Untreated	$1,000,000,000	500,000	$2,000	$500,000,000	100,000	$4,000	$600,000,000	400,000	$1,000
CC-PUF	-$82,521,415	-139,369	$465	-$42,809,959	-9,811	$11	-$39,711,456	-129,558	$421
AL-PUF	$6,304,275	-546	$13	$4,909,217	389	$28	$1,395,058	-935	$6

NOTE: For disclosure concerns, all entries in the tables have been rounded to only one significant digit for untreated data or the 15% sample. CC-PUF and AL-PUF results are shown as differences to illustrate the comparison without sharing actual estimates.

Table 3.1 Diabetes Counts for 2010 Medicare Beneficiaries (Domain: Comorbidty, M , Nondual eligible, Part A)

	IHD	CHF	CAN
Untreated	20,000	10,000	1,000
CC-PUF	-1,998	-1,105	-525
AL-PUF	56	23	-2

Table 3.2 Diabetes Counts for 2010 Medicare Beneficiaries (Domain: Comorbidty, M , Nondual eligible, Part A)

	IHD	CHF	CAN
Untreated	50,000	20,000	6,000
CC-PUF	-3,971	-1,761	-1,097
AL-PUF	-119	67	-36

Table 3.3 Diabetes Counts for 2010 Medicare Beneficiaries (Domain: Comorbidty, M , Nondual eligible, Part A)

	IHD	CHF	CAN
Untreated	60,000	20,000	10,000
CC-PUF	-4,299	-1,809	-1,408
AL-PUF	266	55	16

Table 3.4 Diabetes Counts for 2010 Medicare Beneficiaries (Domain: Comorbidty, M , Nondual eligible, Part A)

	IHD	CHF	CAN
Untreated	50,000	20,000	10,000
CC-PUF	-3,638	-1,729	-1,805
AL-PUF	-438	-285	-165

Table 3.5 Diabetes Counts for 2010 Medicare Beneficiaries (Domain: Comorbidty, M , Nondual eligible, Part A)

	IHD	CHF	CAN
Untreated	40,000	20,000	9,000
CC-PUF	-2,796	-1,564	-1,968
AL-PUF	-28	-146	-41

Table 3.6 Diabetes Counts for 2010 Medicare Beneficiaries (Domain: Comorbidty, M , Nondual eligible, Part A)

	IHD	CHF	CAN
Untreated	30,000	20,000	6,000
CC-PUF	-1,922	-1,314	-1,896
AL-PUF	-217	-159	-64

NOTE: For disclosure concerns, all entries in the tables have been rounded to only one significant digit for untreated data or the 15% sample. CC-PUF and AL-PUF results are shown as differences to illustrate the comparison without sharing actual estimates.

Table 4: Calibration Controls for Subsamples in AL-PUF

Variable	Levels
Age * sex	42
Age * race	105
Age * dual status	42
cc_alzhdmta	2
cc_chf	2
cc_chrnkidn	2
cc_copd	2
cc_depressn	2
cc_diabetes	2
cc_ischmcht	2
cc_osteoprs	2
cc_ra_oa	2
cc_strketia	2
cc_cancer	2
cc_2_or_more	2
TOTAL control totals	213

Table 5: Disclosure Risk Measures based on 500 Simulations

(a) Median δ over MGs

Event	$\tau = .05$		$\tau = .10$		$\tau = .20$	
	(.9,.9,.6)	(.9,.9,.2)	(.9,.9,.6)	(.9,.9,.2)	(.9,.9,.6)	(.9,.9,.2)
AB	.296	.296	.390	.390	.950	.950
DE	.048	.000	.188	.066	.536	.240
ABE	.058	.000	.126	.052	.570	.256
ABCDEF	.016	.000	.082	.000	.392	.158

(b) Third Quartile δ over MGs

Event	$\tau = .05$		$\tau = .10$		$\tau = .20$	
	(.9,.9,.6)	(.9,.9,.2)	(.9,.9,.6)	(.9,.9,.2)	(.9,.9,.6)	(.9,.9,.2)
AB	.348	.348	.462	.462	.960	.960
DE	.068	.004	.204	.110	.566	.254
ABE	.072	.002	.142	.076	.598	.270
ABCDEF	.024	.000	.094	.048	.422	.196

Table 6: Information Loss Measures based on 500 Simulations

(a) Median ε over Domains

Outcome Variable	$\tau = .05$		$\tau = .10$		$\tau = .20$	
	(.9,.9,.6)	(.9,.9,.2)	(.9,.9,.6)	(.9,.9,.2)	(.9,.9,.6)	(.9,.9,.2)
Total Count	.000	.011	.000	.000	.000	.000
Total Expenditure	.055	.331	.000	.048	.000	.000

(b) Third Quartile ε over Domains

Outcome Variable	$\tau = .05$		$\tau = .10$		$\tau = .20$	
	(.9,.9,.6)	(.9,.9,.2)	(.9,.9,.6)	(.9,.9,.2)	(.9,.9,.6)	(.9,.9,.2)
Total Count	.069	.259	.000	.030	.000	.000
Total Expenditure	.244	.540	.010	.227	.000	.018