Aggregate Level PUF as a New Alternative to the Traditional Unit Level PUF for Improving Analytic Utility and Data Confidentiality

A.C. Singh and J. M. Borton
NORC at the University of Chicago, Chicago, IL 60603
singh-avi@norc.org, borton-joshua@norc.org

Abstract

Creating a unit level public use file (PUF) with a rich set of analytic variables and high analytic utility has become a difficult problem due to increasing potential for availability of unit level information in public domains that could be used for matching purposes. Besides, unit level information is prone to possibly false perception of information disclosure about a target of interest which may be difficult to refute. The problem considered here arose in the context of CMS Medicare claims data where unit level corresponds to beneficiaries. To get around this problem, we propose a new approach of aggregate level PUF (or AL-PUF) where we modify the data structure by changing the unit of observation from beneficiaries to a small aggregate (termed micro-group or MG) signifying a group of beneficiaries having a common profile with respect to geo-demographics and prescription drug enrolment. For analytic utility, MG sizes should not be too large in order to make them as close as possible to the unit level; i.e., as building blocks, and for this reason larger MGs could be subdivided using additional outcome variables such as total number of claims and cost for each beneficiary.

The basic idea of MG structure and small MG sizes is motivated from the commonly used aggregate level modeling as an alternative to unit level modeling for small area estimation. In considerations of data confidentiality, however, MG sizes should not be too small either (e.g., not below 10) depending on the level of risk tolerance. Having MGs as building blocks goes a long way in reducing disclosure risk because there is no beneficiary level information. To obtain true totals for various domains, it is sufficient to have only averages (termed micro-means or MMs which are common for all beneficiaries in the MG) of outcome variables for each MG along with MG counts; i.e., weighted up MG sizes. However, for MGs containing single beneficiaries in analytic profiles defining the domains of interest, actual beneficiary values of outcome variables could be disclosed by MG totals.

To mitigate the above disclosure problem, two nested subsamples of the full sample are defined; the larger one for computing MMs for categorical outcome variables or proportions of beneficiaries belonging to analytic profiles for each MG, and the smaller one for MG counts, while the full sample is used to obtain MMs for continuous outcome variables. Subsampling provides unbiased total estimates as well as justification for using two phase sampling results for precision estimation. There is some information loss due to subsampling but it can be minimized by suitably choosing subsampling rates. For increased precision, sampling weights from subsamples are calibrated to the original full sample estimates for key analytic variables. In terms of modeling with AL-PUF data, it is observed that there might be need of instrumental variables (which might be available from a previous or separate independent sample) to avoid bias due to measurement errors

because both dependent variable and independent variables or covariates at the MG level in the form of estimated MMs from the full sample make the model error correlated with the covariates unless the full sample is a census. However, there is no such problem with descriptive inference. Measure sof analytic utility and confidentiality of the proposed method of AL-PUF are illustrated for a 5% sample of the 2008 Medicare Inpatient Claims data.

Key Words: Aggregate Level Analysis; AL-PUF; Calibration; Micro-Group; Micro-Mean; Subsampling; Unit Level Analysis.

1. Introduction

The problem considered in this paper arose in the context of creating unit-level public use files (PUFs) for CMS Medicare claims data where unit level corresponds to the beneficiary level microdata. For a nonsynthetic PUF creation, methods of perturbation and suppression for disclosure treatment are used in a controlled fashion so that information loss and disclosure risk can be kept at or below reasonable tolerance levels; see for example the method of GenMASSC (Singh, 2009). Under this method, uncertainty about the identity and presence of a beneficiary is introduced by random substitution and subsampling. For high analytic utility of the inpatient claims data, we need a rich set of analytic variables in PUF such as age, gender, state, and prescription drug enrolment from the beneficiary summary file, and diagnosis, treatment, cost, payment, utilization, and duration between health episodes. However, with any probabilistic treatment at the unit level, it is difficult to protect against possibly falsely perceived identity disclosure of a beneficiary (which may seem to lead to attribute disclosure as well) if the intruder believes the beneficiary's profile in terms of a set of analytic variable values to be similar to the target. Moreover, with an ever increasing potential of personal health information being available now or in future from various sources that could serve as matching files for identifying variables assumed to be known to the intruder, it makes it harder to refute such claims.

The above concern about the problem of providing adequate data confidentiality led to a compromise solution of a treated controlled use file requiring a less stringent authorization of the data use agreement for access to the treated data housed in a secure environment as well as requiring only a simplified disclosure review before the analysis results could be exported; see Borton et al. (2011). Both user authorization and disclosure review processes can be made simpler and faster due to prior disclosure treatment of the raw microdata. The usual alternative of synthetic PUFs for the claims data is also not feasible because of the difficulty in joint parametric modeling of a large number of analytic variables or in joint nonparametric modeling via empirical distributions. Besides, the problem of perceived disclosure risk continues to persist even for synthetic PUFs.

The dissemination model of treated controlled use file instead of the usual PUF is in line with the important paper of Gomatnam et al. (2005) who argue that the future of usual PUFs (synthetic or nonsynthetic) is rather limited due to the need in practice of a rich set of variables for high analytic utility which does not seem possible while maintaining high data confidentiality for reasons mentioned above. They advocate, instead, for remote analysis servers where the analyst would have only an indirect access to the microdata through web-queries while the raw data is housed in a secure environment. The idea of remote analysis servers has been around for over 30 years, but recently the concern about

PUFs has led to a rejuvenation of research; e.g., see Singh et al. (2012) for a query-based PUF and other references contained there-in.

Despite the promising future of remote analysis servers, there is still need for an inexpensive option of PUF-type data for mass users to gain familiarity with the data, formulate the problem and perform initial analyses before submitting queries through remote analysis servers for final analysis if deemed necessary. In this paper, we propose such an alternative termed aggregate level PUF (AL-PUF for short) which provides a new type of PUF with high analytic utility and data confidentiality. The distinguishing feature of AL-PUF is that unlike usual PUFs which are at the unit or micro level, it is at a small aggregate level termed microgroup or MG. In the case of medicare claims data, the aggregate level signifies a group of beneficiaries having a common profile with respect to geo-demographics and prescription drug enrolment and possibly cross-classified further by total cost and number of claims. The problem of perceived disclosure risk even after sufficient disclosure treatment in unit level PUFs is considerably reduced as we move away from the beneficiary level profile to a beneficiary group level profile. For analytic utility, MG sizes should not be too large in order to make them as close as possible to the unit level; i.e., as building blocks, and for this reason larger MGs are subdivided using additional analytic variables if necessary. The basic idea of MG structure and small MG sizes is motivated from the commonly used aggregate level modeling as an alternative to unit level modeling for small area estimation. For data confidentiality, however, MG size should not be too small either (e.g., not below 10) depending on the level of risk tolerance.

In AL-PUF, having MGs as building blocks goes a long way in reducing disclosure risk because there is no beneficiary level information that is released. To obtain true totals for various domains defined by geo-demographics, diagnosis and treatment, it is sufficient to have only averages (termed micro-means or MMs which are common for all beneficiaries in the MG) of outcome variables for each MG along with MG counts; i.e., weighted up MG sizes. However, for MGs containing single beneficiaries in analytic profiles defining the domains of interest, actual beneficiary values of outcome variables could be disclosed by MG totals. To mitigate this problem, two nested subsamples of the full sample (s_1) are defined; the larger one s_2 (subsample of s_1) for computing MMs or proportions for categorical outcome variables defined by beneficiaries belonging to analytic profiles for each MG, and the smaller one s_3 (subsample of s_2) for MG counts, while the full sample s_1 is used to obtain MMs for continuous outcome variables.

In the absence of any subsampling in AL-PUF, estimates of descriptive parameters such as analysis domain totals from unit level data can be obtained as sums of products of MG counts and MMs which match exactly with the original estimates. However, even if there is no subsampling, there is loss in precision in estimates of model parameters with aggregate level modeling because aggregate level predictors or covariates do not have as much discrimination power as unit level predictors. This is analogous to the use of grouped frequency distribution for estimating moments incurring some loss of efficiency compared to estimates from the ungrouped frequency distribution. Nevertheless, AL-PUF preserves the data integrity better than unit level PUFs by using only subsampling and avoiding distortions due to perturbation and suppression. It may be noted that although aggregate level models are commonly used in small area estimation (an area of great demand in practice; see National Research Council (2000) report on poverty estimation), it is for a different reason due to lack of availability of unit-level predictors and not for data confidentiality reasons.

It is important to note that use of subsampling mentioned above is desirable as it provides unbiased total estimates as well as justifies use of two phase sampling results for precision estimation. There is some information loss due to subsampling but it can be reduced by suitably choosing subsampling rates. For increased precision, sampling weights from subsamples are calibrated to the original full sample estimates for key analytic variables. In terms of modeling with AL-PUF data, it is observed that there would be need of instrumental variables (which might be available from a previous or separate independent sample) to avoid bias due to measurement errors because both dependent and independent variables at the MG level in the form of estimated MMs from the full sample make the model error correlated with the covariates unless the full sample is a census. However, there is no such problem with descriptive inference. It may be remarked that if the microdata are at two levels such as beneficiaries and claims within beneficiaries in the case of medicare (or households and individuals within households in a population survey), separate AL-PUFs can be created at the two levels but employing the same nested subsamples; for instance, MGs can be defined in terms of groups of beneficiaries for the beneficiary level data and in terms of claims from the same groups of beneficiaries for claim level data.

The organization of this paper is as follows. Section 2 provides a heuristic motivation of the proposed method of AL-PUF followed by a stepwise description in Section 3 using the Medicare Claims data as an example. In Section 4, we consider properties of AL-PUF in terms of analytic utility and confidentiality for a given set of subsampling rates using a small simulation study from the 2008 Medicare Inpatient Claims data. An example of the analysis of AL-PUF data in terms of point and variance estimates of descriptive parameters (means, totals, and ratios) and model parameters is discussed in Section 5. In the modeling context, we consider how with two AL-PUFs created from independent samples, one dataset can be used to provide instrumental variables for the other dataset for fitting models. However, for descriptive inference, only one dataset may be used. The last Section 6 contains concluding remarks.

2. Heuristic Motivation of the Proposed Method of AL-PUF

To overcome the difficulty in providing high data confidentiality in the presence of a rich set of analytic variables in PUF, it is clear that some reasonable compromises need to be made. By moving away from unit-level to aggregate-level data, the major problem of possibly false perception of disclosure of individual records essentially disappears by construction as long as the number of observations in each aggregate is not too small. With this in mind, we create aggregates (or micro-groups denoted by MG) of individual records such that eaxh MG size is around 20 (for example) with respect to the full sample or the whole population, thus easily satisfying the rule of 11 provided in CMS data dissemination guidelines. MGs form a partition of the beneficiary population for a given year and can be constructed in terms of basic beneficiary profiles defined by a cross-classification of age, gender, state, beneficiary enrolment, and if necessary, cost and number of claims. Collapsing of MGs may sometimes be necessary to meet the minimum MG size restriction.

To fix ideas, it would be useful to contrast the structure and contents of a unit level with an aggregate level dataset. Table 1 presents unit level data with rows corresponding to different beneficiaries but grouped together in MGs to facilitate comparisons with its aggregate level version in Table 2. Under column 1, he first subscript g in the beneficiary

identification (ID0 is the MG ID and the second subscript *k* is the beneficiary number within MG. The second column shows the sampling weight for each beneficiary based on the original full sample s_1. The third column represents beneficiary's basic profile in terms of binary auxiliary variables $(x_{g1}, x_{g2}, \ldots)$ (or the row vector $\boldsymbol{x}_g$) defining the *g*th MG. The *x*-variables are simply category indicators of variables age, gender, state, enrolment etc. whose cross-classified categories define MGs. For each MG *g* and the beneficiary *k* within MG, the fourth column shows various claim-level analytic profiles (*f*) of interest where *f* varies from 1 to *F*. In the case of Inpatient Claims data, analytic profiles are typically defined by diagnosis and treatment. For each analytic profile, outcome variables corresponding to the *k*th beneficiary within the *g*th MG could be categorical ($\tilde{y}_{gk(f)}$—indicating presence or absence of the profile *f*) or continuous ($y^{(i)}_{gk(f)}$; *i*=1, 2,..) where the *i*th variable denotes length of stay (LOS), cost, or payment, for example, as shown in the last two columns.

In Table 2 representing AL-PUF, all the entries are at the MG level. For the *g*th MG, we have the MG profile ($\boldsymbol{x}_g$), estimated MG count $\widehat{N}_{g(s_3)}$ of beneficiaries, based on a subsample s_3 nested within a larger subsample s_2 of the original full sample s_1, and a set of MG level MMs--proportions $\hat{P}_{g(f,s_2)}$ based on the subsample s_2, and averages $\hat{A}_{g(f,s_1),y^{(i)}}$ based on the full sample s_1 corresponding respectively to categorical and continuous outcome variables for each profile *f*. Note that estimated MG counts $\widehat{N}_{g(s_3)}$, and estimated MMs $\hat{P}_{g(f,s_2)}$ $\hat{A}_{g(f,s_1),y^{(i)}}$ use samples s_3, s_2 and $s_1$respectively and corresponding weights $w_{gk(s_3)}$, $w_{gk(s_2)}$, and $w_{gk(s_1)}$ although weights for subsamples s_3, s_2 are not shown in Table 1. All weights are calibrated to control totals for key analytic variables from the full sample which itself is calibrated first to a key set of known control totals from the 100% claims data. Need for subsampling for protecting confidentiality in AL-PUF is explained below We note that for AL-PUF, MGs serve as building blocks for computing estimates for a variety of analysis domains defined by variables coming from different types of claims data such as inpatient, outpatient, carrier, prescription drug event, skilled nursing facility, home health agency, hospice, and durable medical equipment. MGs are defined such that each beneficiary is assigned to a unique MG. It is for this reason, claims data should not be used for defining MGs because a beneficiary may have different claim types and different claims within a claim type. In any application of AL-PUF, MGs are formed in advance and users are not free to define their own MGs. This restriction is not serious since choice of MGs is based on common analytic needs and is at low levels of aggregation. Incidentally, the above restriction is somewhat analogous to pre-specified coarsening of analytic variables in creating usual PUFs.

Next we consider the need of nested subsampling for AL-PUF in order to create uncertainty in estimated totals for disclosure safety. It has to do with the problem of possible disclosure risk due to rare analytic profiles when MG counts and MMs are computed from the same data set (such as the full sample s_1) which could be the 5% sample of Medicare Claims data commonly used for analysis. By multiplying MMs by the MG count, one could easily obtain the numerator (because MG count is in the denominator of MMs), and if the analytic profile is rare, numerators are likely to be based on a single beneficiary and their values would simply be the product of the sampling weight and values of outcome variables for the beneficiary in the rare profile. Thus the sensitive values might be at risk of disclosure. To avoid this problem, we propose using a subsample such that the full sample is used to estimate MMs for each MG and the

subsample to estimate MG counts. It is easily seen that in estimating total for any outcome variable, the product of MG count and MM will no longer yield the numerator because MG counts (in numerator and denominator) do not cancel out.

There is still one more problem. The MM for the variable $\tilde{y}_{gk(f)}$ indicating presence or absence of a beneficiary in an analytic profile has the same denominator (i.e., the estimated MG count based on the full sample s_1) as the other MMs for continuous outcome variables $y^{(i)}_{gk(f)}$. Therefore, by dividing any one of the MMs of $y^{(i)}_{gk(f)}$ by MM of $\tilde{y}_{gk(f)}$, we can recover the ratio of the numerators of MMs. Now, if the analytic profile is rare, again the value of a sensitive variable might be at risk of disclosure. To counter this problem, we propose to use two nested subsamples s_3 and s_2 of s_1 where s_3 is a subsample of s_2, such that for each analytic profile, s_1 is used for estimating all MMs $\hat{A}_{g(f,\,s_1),y^{(i)}}$, s_2 is used for estimating the MM $\hat{P}_{g(f,\,s_2)}$, and s_3 is used for estimating the MG count $\widehat{N}_{g(s_3)}$ as shown in Table 2. In terms of estimation efficiency, it implies that the resulting domain estimates will be more efficient than just using the smallest s_3 sample, but less efficient than using the largest s_1 sample for all components—MG counts and MMs.

So far we were mainly concerned with estimating domain totals; i.e., descriptive parameters. However, for fitting models for a dependent variable defined by MM of an outcome variable at the MG level, we can use covariates defined by MMs for related analytic variables. However, this will lead to biased estimates as in the case of models with measurement errors in covariates because covariates are estimated MMs which are correlated with model errors as both dependent and independent variables are based on the same sample dataset. To remedy this problem, we propose using two independent sets of three nested samples (s_3, s_2, s_1) and (s_3^*, s_2^*, s_1^*) so that MMs from the second set serve as instrumental variables for fitting models based on the first set. In practice, MGs may have to be grouped to form small domains before model fitting to obtain a stable error covariance structure at the domain level as in aggregate level small area modeling.

3. Description of the Proposed Method for Creating AL-PUF

Based on the brief description of AL-PUF in the previous section, it is observed that once the MG structure along with MG counts for creating AL-PUF from a dataset is specified, it may be more convenient for analysts if the AL-PUF is expressed in a different order from the order in Table 2. Specifically, consider subtables consisting of MGs with MMs where each subtable corresponds to a given analytic profile (*f*) as shown in Table 3. This is in contrast to using subtables of analytic profiles with MMs where each subtable corresponds to a given MG shown within a given row of MG, and together forming the very large complete Table 2. It follows that in practice a library of analytic profiles with unique IDs can be constructed and updated over time as more years of data and new requests arrive for analysis domains defined by basic beneficiary level profiles and cliam level analytic profiles. The number of such subtables is likely to be rather large because the number of possible analytic profiles of interest could be very large. However, in practice, this may not be of much concern in view of ever increasing memory and storage capabilities of modern computers. Besides, having information on analytic profiles in AL-PUF avoids the tedious task for analysts to extract it directly from the complex raw claims data.

It may be noted that after the initial construction of AL-PUF, it remains quite flexible with regard to updating it with more subtables for new analytic profiles over time without any impact on earlier ones. This feature renders AL-PUF user-friendly with high analytic utility. Here, it may be illustrative to list some examples of clinical or analytic profiles (AP) from inpatient claims data: AP_1: Diabetes in Yr 1; AP_2: Diabetes in Yr 2; AP_3: Diabetes in Yr 3; AP_4: CAD in Yr 1; AP_5: CAD in Yr 2; AP_6: CAD in Yr 3; AP_7: Bypass in Yr 1; AP_8: Bypass in Yr 2; AP_9: Bypass in Yr 3; $AP_{2,5}$: Diabetes in Yr 2 and CAD in Yr 2; $AP_{1,2,5,8}$: Diabetes in Yr 1 and Yr 2, CAD in Yr 2, and Bypass in Yr 2, and so on. Analytic profiles may have detailed ICD-9 codes for diagnoses and procedures and there is no suppression of any analytic profile in AL-PUF unless some reasonable broad requirements are imposed; e.g., at least three MGs with nonzero values of outcome variables for each analytic profile may be required.

A detailed stepwise description of AL-PUF now follows.

Step I. Partition the full sample (s_1) of beneficiaries into small subgroups or micro-groups (MGs) of size around 20 by cross-classifying demography, geography, and enrolment. Split large MGs into smaller ones by further cross-classifying with other outcome variables such as cost and number of claims which are not specific to a particular claim type. Collapse MGs if necessary to satisfy the minimum sample size.

Step II. Obtain suitable control totals for calibration from a larger dataset such as the complete Medicare claims data. Choose a sample (s_2) as a subsample of $s_1$and then a sample (s_3) as a subsample of s_2. Perform weight calibration for all the three samples (s_3, s_2, s_1) to the same set of calibration controls.

Step III. Estimate MG counts using calibrated weights from the s_3 sample**.** Next, for each analytic profile of interest from a given claims file, define MMs ($\hat{A}_{g(f,\,s_1),y^{(i)}}$) using calibrated weights for the s_1 sample except using s_2 weights for the proportion ($\hat{P}_{g(f,\,s_2)}$) of beneficiaries in the analytic profile. Thus the variables MG count and MM can be populated for each claims data file corresponding to all MGs for which MG count is not zero in the s_3 sample; see Table 3. Also add columns of MMs for squares and cross-products of outcome variables needed for variance estimation as explained in Section 5.

Step IV. Depending on the analysis, extract suitable subsets of data on analytic profiles of interest from profile-specific files using profile IDs; each profile-specific file gives rise to a subtable of the form Table 3. For instrumental variables required in modeling, repeat the above process with an independent sample and its subsamples (s_3^*, s_2^*, s_1^*).

4. Measures of Analytic Utility and Confidentiality of AL-PUF

If the full sample s_1 were used without subsampling, there would have been no loss of information for estimating descriptive parameters and the analytic utility of AL-PUF same as that of the original data. However, subsampling is needed for protecting confidentiality as explained in Section 2. In this process of trade-off, analytic utility is affected. With suitable choices of subsampling rates for s_2 out of s_1, and s_3 out of s_2, it is possible to have both high confidentiality and analytic utility. With Medicare Claims data, typically 5% files based on simple random samples of the whole administrative dataset are made available to researchers. It seems natural then to consider three non-

overlapping and (approximately) independent 5% samples and treat the combined 15% sample as s_1, a combination of two 5% samples as s_2, and one of the 5% samples as s_3. With respect to the full sample, this amounts to subsampling rates of 50% for s_2 and 33.33% for s_3. For protecting confidentiality, we need to check whether there is sufficient fluctuation in estimated MG counts $\widehat{N}_{g(s_3)}$ in relation to $\widehat{N}_{g(s_2)}$, and $\widehat{N}_{g(s_2)}$ in relation to $\widehat{N}_{g(s_1)}$. This can be measured through simulations of nested subsamples to compute mean absolute relative error (MARE) for each MG for each of the two subsamples for a given full sample. It turns out that several versions of the AL-PUF framework MG counts for a given set of MGs based on the original full sample can be easily created by repeated subsampling (typically stratified simple random) of the full sample if sampling weight for subsamples are not calibrated. Calibration step can be omitted for simulations because it is not likely to affect much the variability in MG counts. However, it is needed for the final AL-PUF based on one set of subsamples for improved precision in analysis domain estimates.

We now consider confidentiality (or inverse disclosure risk) measures. For the *m*th simulation, *m*= 1, …,*M*, we define

$$MARE\left(\widehat{N}_{g(s_2)}\right) = \frac{1}{M}\sum_{m=1}^{M} \frac{\left|\widehat{N}_{g(s_2)}^{(m)} - \widehat{N}_{g(s_1)}\right|}{\widehat{N}_{g(s_1)}} \ , \qquad (1a)$$

and for each subsubsample $s_3^{(m)}$ of the subsample $s_2^{(m)}$ over *M* simulations, define

$$MARE\left(\widehat{N}_{g(s_3)}\right) = \frac{1}{M}\sum_{m=1}^{M} \frac{\left|\widehat{N}_{g(s_3)}^{(m)} - \widehat{N}_{g(s_2)}^{(m)}\right|}{\widehat{N}_{g(s_2)}^{(m)}} \ . \qquad (1b)$$

From the above MAREs for each MG, two sets of confidentiality measures can be computed based on quantiles of $MARE\left(\widehat{N}_{g(s_2)}\right)$, and $MARE\left(\widehat{N}_{g(s_3)}\right)$. A rule of thumb in practice might be to choose subsampling rates such that these confidentiality measures are not below 20%.

For measures of information loss (or inverse analytic utility), we can again use simulations to define MARE of analysis domain level count or total estimates where domains are defined by basic beneficiary and claim level analytic profiles. Thus each analysis domain *d* (varying from 1 to *D*) can be defined as a universe U_d consisting of a group of g's crossed by a group of f's. Now for the categorical outcome variable $\tilde{y}_{gk(f)}$, consider the estimated count $\widehat{N}_d$ for domain *d* given by

$$\widehat{N}_d = \sum_{f=1}^{F}\sum_{g=1}^{G} \widehat{N}_{g(s_3)}\, \hat{P}_{g(f,\, s_2)} 1_{(g,f)\in U_d} \qquad (2a)$$

where $1_{(g,f)\in U_d}$ takes the value of 1 if (g, f)belong to the domain *d* and 0 otherwise. The true domain count N_d^* based on the sample $s_1$is given by

$$N_d^* = \sum_{f=1}^{F}\sum_{g=1}^{G} \widehat{N}_{g(s_1)}\, \hat{P}_{g(f,\, s_1)} 1_{(g,f)\in U_d} \qquad (2a)$$

where both $\widehat{N}_{g(s_1)}$ and $\hat{P}_{g(f,\,s_1)}$ are based on the full sample s_1. Similarly, the estimated total for a continuous variable $y^{(i)}_{gk(f)}$ for domain *d* is given by

$$\hat{T}_{d,y^{(i)}} = \sum_{f=1}^{F}\sum_{g=1}^{G} \widehat{N}_{g(s_3)}\, \hat{A}_{g(f,\,s_1),y^{(i)}} 1_{(g,f)\in U_d} \qquad (3a)$$

and the corresponding true total based on the full sample is given by

$$\begin{aligned} T^{*}_{d,y^{(i)}} &= \sum_{f=1}^{F}\sum_{g=1}^{G} \widehat{N}_{g(s_1)}\, \hat{A}_{g(f,\,s_1),y^{(i)}} 1_{(g,f)\in U_d} \\ &= \sum_{f=1}^{F}\sum_{g=1}^{G} T^{*}_{g(f,\,s_1),y^{(i)}}\, 1_{(g,f)\in U_d}. \end{aligned} \qquad (3b)$$

Now measures of information loss over a set of domains can be defined as quantiles of MARE of $\widehat{N}_d$ relative to N^{*}_d, and MARE of $\hat{T}_{d,y^{(i)}}$ relative to $T^{*}_{d,y^{(i)}}$ from *M* simulations as before. Note that for expediency at the design stage of AL-PUF, the subsampling weights for each simulation are not calibrated which would tend to reduce the precision of domain estimates; i.e., observed MARE are likely to appear greater than they really are with calibrated weights. This gives rise to a conservative rule of thumb which we could set at 20% for the upper limit on MARE for domain estimates but for most estimates preferably under 15%.

To illustrate computation of above measures of disclosure risk and information loss, we considered a limited simulation study with *M*=1000 subsamples from a 2008 5% sample of Inpatient Claims data with subsampling rates of 40% for s_2 (i.e., a 2% sample of the claims data) and 20% for s_3(i.e., a 1% sample of the claims data). Stratified simple random sampling was used for subsampling within strata of the full sample defined by broad categories of age, race/ethnicity, state and gender. Sampling weights were not calibrated for reasons mentioned above. For measuring disclosure risk, MARE was computed somewhat differently from the formulas 1(a) and (b). MARE of $\widehat{N}^{(m)}_{g(s_3)}$ was calculated relative to $\widehat{N}_{g(s_1)}$ and not $\widehat{N}^{(m)}_{g(s_2)}$, and instead of $MARE\big(\widehat{N}_{g(s_2)}\big)$, we computed MARE $(1/\hat{P}_{g(f,\,s_2)})$ where the analytic profile f was taken as cardiac bypass with and without MCC. The simplified measures of risk proposed here were not developed before the simulation study was conducted. For measuring information loss, MARE was computed for two outcome variables: basic demographic domain counts, and analytic domain counts defined by demographic and analytic profile of cardiac bypass with and without MCC, where demographic domains were defined by gender by age categories in terms of year of birth (1922 or before, 1923-1927, 1928-1932, 1933-1937, 1938-1942, 1943 or later); twelve in all. Table 4 shows various measures. The number of MGs formed was 95664 varying in size from 25 to 37, and the number of MGs with at least one claim for the analytic profile considered was 2213. It is seen that for the particular choice of subsampling rates, confidentiality measures for both subsamples are reasonably large (above 20% for all MGs), and measures of information loss for both domain types are reasonably small (below 20% for all domains).

5. Analysis Examples with AL-PUF (Medicare Inpatient Claims Data)

We describe a few simple examples of descriptive inference from an AL-PUF data for Medicare Inpatient Claims. Consider three analysis domains defined by diagnosis and treatment profiles only; i.e., without being crossed by basic beneficiary geo-demographic profiles:

d_1: Beneficiaries diagnosed with coronary artery disease (CAD),
d_2: Beneficiaries diagnosed with coronary artery disease and received bypass, and
d_3: Beneficiaries with coronary artery disease and received angioplasty.

The parameters of interest are N_{d_1}, N_{d_2}, N_{d_3}, and their ratios N_{d_2}/N_{d_1}; i.e., proportion of beneficiaries diagnosed with CAD that received bypass, and N_{d_3}/N_{d_1}; i.e., the proportion of beneficiaries diagnosed with CAD that received angioplasty. The point estimates of total counts are:

$\widehat{N}_{d_1}$=Estimated total number of beneficiaries diagnosed with CAD
= Sum over all MGs of MM (proportion of MG with CAD based on s_2) times MG Count (based on s_2)

$$= \sum_{f=1}^{F}\sum_{g=1}^{G} \widehat{N}_{g(s_3)}\, \hat{P}_{g(f,\,s_2)} 1_{f=CAD}$$

Similarly,

$$\widehat{N}_{d_2} = \sum_{f=1}^{F}\sum_{g=1}^{G} \widehat{N}_{g(s_3)}\, \hat{P}_{g(f,\,s_2)} 1_{f=CAD\ and\ Bypass}\ ,$$

and $$\widehat{N}_{d_3} = \sum_{f=1}^{F}\sum_{g=1}^{G} \widehat{N}_{g(s_3)}\, \hat{P}_{g(f,\,s_2)} 1_{f=CAD\ and\ Angioplasty}.$$

All the above estimates are unbiased using standard arguments of two-phase sampling. If interested in average cost per beneficiary, the parameters of interest become $T_{d_2,y}/N_{d_1}$, and $T_{d_3,y}/N_{d_1}$where y is the cost variable. These parameters can be estimated in an analogous manner with $\hat{P}_{g(f,\,s_2)}$ replaced by $\hat{A}_{g(f,\,s_1),y}$ in the above formulas.

For variance estimation, two-phase results are applicable after linearization of ratio estimates. In particular, as shown in Appendix I, for simple random samples at both phases, usual variance estimate can be approximated quite well by aggregate level data (MMs and MG counts) in AL-PUF as long as number of MGs is large enough which holds for large samples. If the data is from a complex multistage design, usual with replacement primary sampling unit (PSU) formulas can be used provided MGs are formed within PSUs, and PSUs are treated as strata for second phase subsampling. This ensures PSU level estimates from AL-PUF to be unbiased and independent across PSUs. However, some PSUs may need to be collapsed in order to have a sufficient number of individuals for forming MGs. Moreover, PSU IDs for each MG need to be added to AL-PUF for above variance estimation.

For model fitting under analytic inference based on a given AL-PUF with some MMs as auxiliary variables and other MMs as dependent variables, we will need a second AL-PUF to provide corresponding MMs for auxiliary variables as instrumental variables for unbiased parameter estimation. The method of quasi-likelihood under a given error

covariance structure can be used for this purpose. For example, a log-linear model for total in-patient expenditure as a function of demographic and chronic conditions can be fitted using the aggregate level data of AL-PUF. Aggregate level model diagnostics can also be performed as in aggregate level approach to small area estimation.

6. Concluding Remarks

In this paper we exploit the connection between aggregate level modeling and unit level modeling when unit level auxiliary information is not available or cannot be released due to confidentiality reasons. The basic idea is rooted in commonly used aggregate level modeling approach for small area estimation. It gives rise to a new PUF, termed AL-PUF, which has high analytic utility and data confidentiality compared to traditional unit level PUFs. In using AL-PUF the confidentiality protection is high because with aggregate level, the most difficult problem of protecting against perceived disclosure risk associated with unit level files even after disclosure treatment disappears. It is further controlled by subsampling. The analytic utility remains high for several reasons: one, the MG size is kept small to make it close to the unit level; two, there is no problem of bias due to usual perturbation and suppression needed for unit level PUF because only subsampling is used for introducing uncertainty; three, subsampling rates can be made reasonably high while controlling disclosure risk; and four, there is no problem in adding information about more analytic variables as need arises.

The basic structure of AL-PUF consists of MGs whose formation is based on a large sample such as the combined three 5% samples from Medicare Claims data, and two nested subsamples (such as 10% and 5% for claims data) whose weights are calibrated to the full sample to obtain calibrated MG counts. Next for each analytic variable (defined in general by basic beneficiary profile, clinical or analytic profile, and outcome variables), MMs are computed for each MG and analytic profile. Thus there is no problem in adding more claim-type analytic variables for a given year. All that is needed is to compute MMs for each MG. Similarly, for adding analytic variables across years, we can make MGs common over years by using a particular year (such as 2008) as the reference year, and then new MMs can be easily added. Thus for longitudinal data such as a 3-year CMS claims data, MGs can be used to link data over time even though there may not be a complete overlap of beneficiaries for the same MG between two years due to population changes by birth or death of beneficiaries or other reasons. So MMs from corresponding year-specific claims files for each MG can be linked together to construct a three year AL-PUF.

It follows that AL-PUF is quite flexible and adaptable to including a rich set of analytic variables. Even for sequence of health events over time, suitable analytic profiles can be constructed that capture patterns of interest. For dealing with provider information while protecting their IDs, again suitable analytic profile capturing provider preference by beneficiaries or type of service provided can be constructed. Finally we note that the AL-PUF data structure may, in fact, be preferable to analysts instead of the microdata because there is no need to go through the tedious task of creating a user-friendly analytic summary file from the original raw microdata before performing any analysis. Al-PUF essentially consists of subfiles, each subfile corresponds to a given analytic variable. Instead of users creating their own subsets of data needed for a given analysis, the AL-PUF data producer creates appropriate subfiles. Moreover, with AL-PUF being a public use file, there is no need for any data use agreement for users typically required for access to the raw microdata.

Appendix I

If a multi-stage design is used for the full sample s_1, we could use the usual with replacement PSU assumption to obtain conservative variance estimates of a domain total estimate $\hat{T}_{d,y}$ defined in (3a). We express the total estimator as a sum of PSU-level estimates $t_{j(d,y)}$, *j=1,.., J; J* being the total number of PSUs. An estimate of the variance of $\hat{T}_{d,y}$ is then obtained as

$$\frac{J}{J-1}\sum_{j=1}^{J}(t_{j(d,y)} - \bar{t}_{d,y})^2, \text{ where } \bar{t}_{d,y} = \frac{\sum_{j=1}^{J} t_{j(d,y)}}{J} \qquad \text{(A1)}$$

Note that $t_{j(d,y)}$ involves calibrated weights from both samples s_1, s_3. If the design is simple random sample without replacement, an approximation to the usual two phase variance estimator can be obtained that relies on only aggregate level data as in AL-PUF. More specifically, we have

$$Var\,(\hat{T}_{d,y}) = Var\left(\sum_{f=1}^{F}\sum_{g=1}^{G}\hat{N}_{g(s_3)}\,\hat{A}_{g(f,\,s_1),y}1_{(g,f)\in U_d}\right) = V_2 + V_1 \qquad \text{(A2)}$$

where V_2 is the second phase variance based on the subsample s_3 conditional on the first phase sample s_1, and $V_1$is the first phase variance based on s_1. Observe that the estimator $\hat{T}_{d,y}$ can be expressed as a sum of contributions from each MG; i.e., $\sum_{g=1}^{G}\hat{A}_{g(d,y)}\,\hat{N}_{g(s_3)}$ where $\hat{A}_{g(d,y)}$ is defined as $\sum_{f=1}^{F}\hat{A}_{g(f,\,s_1),y}1_{(g,f)\in U_d}$. It follows that given s_1, $\hat{T}_{d,y}$ is a linear combination of *G* MG-count estimates $\hat{N}_{g(s_3)}$ and so an estimate of V_2 can be easily obtained from the variance-covariance matrix of the *G*-vector of counts $\hat{N}_{g(s_3)}$ under simple random sampling. Next observe that V_1 is the variance of $T^*_{d,y}$ defined in (3b) which can be re-expressed as $\sum_{g=1}^{G}\left(\sum_{f=1}^{F}T^*_{g(f,\,s_1),y}1_{(g,f)\in U_d}\right)$ or $\sum_{g=1}^{G}\sum_{k=1}^{m_g}\sum_{f=1}^{F}y_{gk(f)}1_{(g,f)\in U_d}w_{gk1}$ where $w_{gk1}$is the calibrated weight for sample s_1. Denoting $y_{gk(f)}1_{(g,f)\in U_d}w_{gk1}$ by $z_{gk(f)}$, V_1 under simple random sampling is estimated as

$$\left(1-\frac{n_1}{N}\right)\frac{n_1}{n_1-1}\sum_{g,k,f}\left(z_{gk(f)} - \bar{z}\right)^2 . \qquad \text{(A3)}$$

However, under AL-PUF, $z_{gk(f)}$ are not available but we do have estimates of the total $T^*_{g(d,y)}$ as $\hat{T}_{g(d,y)} = \sum_{g=1}^{G}\hat{A}_{g(d,y)}\,\hat{N}_{g(s_3)}$ which also estimates MG total for $z_{gk(f)}$. If in AL-PUF, we also include the column of y^2w_1as shown in Table 3, estimates of MG totals of $z^2_{gk(f)}$ (to be denoted by $\hat{T}_{g(d,y^2w_1)}$) can be obtained. Now for large *G*, V_1can be estimated by

$$\left(1-\frac{n_1}{N}\right)\frac{n_1}{n_1-1}\left(\sum_g \hat{T}_{g(d,y^2w_1)} - \frac{\hat{T}_{d,y}{}^2}{n_1}\right). \qquad \text{(A4)}$$

It is remarked that the condition of large *G* is required so that $\hat{T}_{d,y}{}^2$ can consistently estimate $T^{*2}_{d,y}$. Also note that in Table 3, the column for the outcome variable for squares are multiplied by w_1—calibrated weight for s_1 so that we automatically get w_1^2 in expressions for $\hat{T}_{g(d,y^2w_1)}$. Similarly, the column for cross-products of *y*-variables is also multiplied by w_1 which makes the appropriate adjustment when finding covariances of

total estimates. Finally, we note that the above formulas can be easily modified if stratified random sampling is used in phase 1 or phase 2.

Disclaimer and Acknowledgments

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, Craig Coelen, Erkan Erdem and Slava Katz of IMPAQ for helpful discussions, and Mike Davern and Susan Hinkins of NORC for useful comments and suggestions. We would also like to thank John Eltinge of BLS for a very useful discussion and Amanda Yu of NORC for suggesting the two analysis examples discussed for the Medicare Claims data application.

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.

Gomatnam, S, Karr, A.F., Reiter, J.P, and Sanil, A.P. (2005). Data Dissemination and Disclosure limitation in a World without Micro-data: A risk-Utility Framework for Remote Access Analytic Servers. *Statistical Science*, Vol 20, 163-177.

National Research Council (2000). *Small- area Estimates of School-Age Children in Poverty: evaluation of Current Methodology*, C.F. Citro and G. Kalton (Eds.), Committee on National Statistics, Washington, DC: National Academy Press.

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 et al. (2012).

Table 1: Unit Level Representation of Medicare Beneficiary and Inpatient Claims Data (Full Sample s_1)

Bene ID	**Sample Weight** $w_{gk(s_1)}$	**Bene Basic Profile in MG** (x_g)	**Claims Analytic Profile (*f*) for each Bene (*gk*)**	**Outcome Variables**	
				Categorical ($\tilde{y}$) Presence or Absence of the profile '*f*'	**Continuous (y): $y^{(1)}$ --LOS, $y^{(2)}$--Cost, $y^{(3)}$--Payment**
11 . $1k$. $1m_1$	$w_{11(s_1)}$. $w_{1k(s_1)}$. $w_{1m_1(s_1)}$	common Profile for all bene in g=1 $(x_{11}, x_{12}, \dots)$	.	.	.
.	.	.	.	.	.
$g1$. gk . gm_g	$w_{g1(s_1)}$. $w_{gk(s_1)}$. $w_{gm_g(s_1)}$	$(x_{g1}, x_{g2}, \dots)$	gk { 1, . , f, . , F	$\tilde{y}_{gk(1)}$. $\tilde{y}_{gk(f)}$. $\tilde{y}_{gk(F)}$	$y^{(1)}_{gk(1)}, y^{(2)}_{gk(1)}, y^{(3)}_{gk(1)}$. $y^{(1)}_{gk(f)}, y^{(2)}_{gk(f)}, y^{(3)}_{gk(f)}$. $y^{(1)}_{gk(F)}, y^{(2)}_{gk(F)}, y^{(3)}_{gk(F)}$
.	.	.	.	.	.
$G1$. Gk . Gm_G	$w_{G1(s_1)}$. $w_{Gk(s_1)}$. $w_{Gm_G(s_1)}$	$(x_{G1}, x_{G2}, \dots)$	.	.	.

Table 2: Aggregate Level Representation of Medicare Beneficiary and Inpatient Claims Data (Full Sample s_1, and Nested Subsamples s_2, s_3)

MG ID (*g*)	**MG or Bene Basic Profile** (x_g)	**MG Count** $\hat{N}_{g(s_3)}$	**Claims Analytic Profile (*f*)**	**Average Outcome Variable (MM)**	
				Proportion of $\tilde{y}$ for *f*, $\hat{P}_{g(f,s_2)}$	**Average of *y* for *f*,** $\hat{A}_{g(f,s_1),y^{(i)}}$ **, for *i*=1, 2, 3**
1	.	.	.	.	.
.	.	.	.	.	.
g	$(x_{1g}, x_{2g}, \dots)$	Using Subsample s_3 $\hat{N}_{g(s_3)}$= $\sum_k w_{gk(s_3)}$	1, . , f, . , F	Using Subsample s_2, $\hat{P}_{g(f,s_2)}$ = $\frac{\sum_k w_{gk(s_2)} \tilde{y}_{gk(f)}}{\hat{N}_{g(s_2)}}$	Using full sample s_1 $\hat{A}_{g(f,s_1),y^{(i)}}$= $\frac{\sum_k w_{gk(s_1)} y^{(i)}_{gk(f)}}{\hat{N}_{g(s_1)}}$
.	.	.	.	.	.
G		.	.	.	.

Table 3: AL-PUF Subtable for a Given Analytic Profile *f*

MG ID	MG Profile	MG Count	MM for $\tilde{y}$	MM for $y^{(i)}$	MM for $\left(y^{(i)}\right)^2 w_1$	MM for $y^{(i)}y^{(j)}w_1$
1	x_1	$\widehat{N}_{1(s_3)}$	$\hat{P}_{1(f,s_2)}$	$\hat{A}_{1(f,s_1),y^{(i)}}$	$\hat{A}_{1(f,s_1),\left(y^{(i)}\right)^2 w_1}$	$\hat{A}_{1(f,s_1),y^{(i)}y^{(j)}w_1}$
.	.	.	.	.	.	.
g	x_g	$\widehat{N}_{g(s_3)}$	$\hat{P}_{g(f,s_2)}$	$\hat{A}_{g(f,s_1),y^{(i)}}$	$\hat{A}_{g(f,s_1),\left(y^{(i)}\right)^2 w_1}$	$\hat{A}_{g(f,s_1),y^{(i)}y^{(j)}w_1}$
.	.	.	.	.	.	.
G	x_G	$\widehat{N}_{G(s_3)}$	$\hat{P}_{G(f,s_2)}$	$\hat{A}_{G(f,s_1),y^{(i)}}$	$\hat{A}_{G(f,s_1),\left(y^{(i)}\right)^2 w_1}$	$\hat{A}_{G(f,s_1),y^{(i)}y^{(j)}w_1}$

Footnote: w_1 denotes the calibrated weight for sample s_1.

Table 4: Simulation-based Measures of Disclosure Risk and Information Loss
(s_1=5%, s_2=2%, s_3=1%; M=1000)

	Minimum	Median	Maximum
Confidentiality (or Inverse Disclosure Risk) Measures			
$MARE\left(\widehat{N}_{g(s_3)}\right)$	24.83%	31.36%	36.03%
MARE $(1/\hat{P}_{g(f,s_2)})$	20.06%	23.01%	57.01%
Information Loss Measures			
MARE $(\widehat{N}_d)$	0.27%	0.34%	0.55%
MARE $(\widehat{N}_{df})$	5.31%	8.30%	18.86%