Query-based PUFs for Disclosure-Safe Remote Analysis from Medicare Claims Micro Data

A.C. Singh, J. M. Borton, M. E. Davern, and Y. Lin
NORC at the University of Chicago, Chicago, IL 60603
singh-avi@norc.org; borton-joshua@norc.org; davern-michael@norc.org; lin-yongheng@norc.org;

Abstract

Remote analysis servers with disclosure-treatment of analysis output from user queries, as an alternative to traditional input disclosure-treated PUFs, form an active research area and are likely to be the future mode of output dissemination from data analysis. The main reason for this is that the preponderance of publically available datasets containing indirect identifiers at present or in future raises concerns about disclosure-safety of traditional PUFs based on static assumptions of intruder knowledge. However, success of remote analysis servers depends on protection from the challenging problem of possible differencing attacks by repeated queries. We describe a new application of a recently proposed method of query-based public use file (or Q-PUF) to Medicare claims data that is not vulnerable to differencing attacks. The reason for this is that in Q-PUF, the user is not allowed to arbitrarily define analysis domains but is required to choose from a checklist of pre-screened variables such that the contributor-count; i.e., the number of observations making contributions to the analysis domains defined by these variables satisfies a minimum threshold. The method is termed PUF, despite being an output treatment, because analogous to the traditional input-treated PUFs, the data producer controls the type and scope of allowed analytic variables. There are four main components of Q-PUF: first, construct a checklist of variables defining analysis domains for which the number of contributors from the data is deemed adequately large to provide reliable estimating functions of corresponding parameters, and hence rendering them automatically disclosure-safe for analysis; second, perform a disclosure audit to choose data-specific adequate confidentiality threshold; third, to provide an interface for users to communicate with the microdata via queries; and fourth, impose additional restrictions specific to the analysis output if needed. The checklist of allowed variables can be updated over time to accommodate new queries as long as disclosure-safety of existing domains is not jeopardized. We provide empirical examples as an illustration of descriptive inference using a working synthetic PUF created from Medicare claims data.

Key Words: Differencing Attacks; Remote Analysis Servers; Input vs. Output Disclosure Treatment; Checklist of Pre-screened Analytic Variables.

1. Introduction

We motivate the general problem of developing a query-based approach to analysis output disclosure treatment as follows. First we note that the basic data framework required for any analysis should be able to support inferential estimation and testing which essentially require computation of estimating functions and corresponding variance estimates; see e.g., Godambe and Thompson (1989, 2009) and McCullagh and Nelder (1989, Chapter 9). An estimating function is a sum of elementary estimating

functions where each elementary estimating function is a function (linear or nonlinear) of data or parameters or both and has mean 0. Thus, given parameters, only aggregate level information from the unit level microdata is needed for computing totals of elementary estimating functions, and so a method of analysis output disclosure treatment needs to be able to ensure that any confidential information about an individual is not released from values of estimating functions. For descriptive inference, such a disclosure-safe data framework for estimating functions is generally adequate for point and variance estimation except for estimating some parameters such as minimum and maximum values in a distribution. Similarly, for analytic or model-based inference, the above framework is also generally adequate for estimating model parameters except for certain parameters in some circumstances, unit-level predictions and residuals for model diagnostics. Therefore, additional output treatment may be needed for disclosure-prone unit-level output or certain parameter estimates such as regression coefficients when R-square value is almost 1 (Lucero et al., 2011). Here the output treatment could simply entail suppression of such information or releasing a lower bound in the case of R^2 or transformation of model residuals to an aggregate level version as discussed in Section 6. For the important problem of model diagnostics using residuals, an alternative approach based on an innovative application of multiple imputation for producing synthetic residuals was proposed by Reiter (2003a).

For a simple illustrative example of the use of the above general framework of estimating functions for descriptive parameters such as population counts and totals for domains (or subpopulations) of interest, consider the Data Entrepreneurs' Synthetic Public Use File (De-SynPUF) for CMS Medicare Claims Data for years 2008-2010 (www.cms.gov) containing about 7.5 million records representing a 5% sample per year. Table 1 is based on an extract of a 50% sample of DE-SynPUF where domains of interest are defined by a cross-classification of age (single years except for bottom and top categories of 50 and under and 90 and over respectively), race, gender and diabetes. Table 1A, in particular, shows observed domain counts and total amounts of in-patient expenditure as well as counts of contributors to expenditure for Hispanic females. The corresponding formulation of estimating functions for estimating domain population counts and totals is considered in Section 5 which is seen to give rise to usual estimates.

The main purpose of this paper is to discuss an application of a query-based method proposed by Singh, Borton, and Crego (2012a) termed query-based PUF (or Q-PUF) to CMS Medicare Claims data. Q-PUF offers a balance between very high data confidentiality and data utility but users are not free to submit queries about arbitrary analysis domains. In particular, Q-PUF puts restrictions on the allowable analysis domains through screening of analytic variables so that the domain size (i.e., the number of contributors to the domain) meets a minimum threshold—denoted D* if the variable is categorical or GD* if it is continuous as described in Section 2. The goal of Q-PUF is to be able to provide descriptive parameter estimates (such as means, totals, ratios, and percentiles but not min and max) along with standard errors, and analytic parameter estimates (such as regression model parameter estimates and predictions subject to certain restrictions) and their standard errors but only for variables in the checklist comprising pre-screened variables. For variables not in the checklist, model-based estimates of domain totals can be produced. It is because of pre-screening of analysis variables that it does not require a DUA (data use agreement) and can be viewed as a PUF although not in the traditional PUF sense because it entails disseminating de-identified analysis output data and not any de-identified input data to be used for subsequent analyses.

2. Domain or Cell Aggregation for Disclosure-Safety

We observe that analysis domains are typically formed by cells or their aggregates from a cross-classified table of a moderate number (7 ± 2 or so) of variables at a time although there could be many more variables in the dataset. Therefore, to ensure disclosure-safety of estimated totals, it is natural to consider the size or the number of contributors in the domain of interest. In other words, the size of the support of estimating functions should be sufficiently large. In the case of population count parameters such as the number of diabetic medicare beneficiaries, this amounts to the specification of a minimum threshold D* (denoting DSTAR--domain size threshold for analysis restrictions) for confidentiality protection of domains (defined by cells or cell aggregates) in tabular data, while in the case of population totals of continuous variables such as the expenditure amount, a threshold for a modified count is needed which was termed GD* (denoting generalized D*) by Singh, Borton, and Crego (2012a). The modified count for each domain is defined by the number of contributors in the data to the domain total. This modified count is not released because it will put individuals with zero contribution at risk of disclosure. In our application, suppose we set in a somewhat ad hoc manner GD* for expenditure at 10 for illustration and D* for diabetes counts at 50. The threshold for diabetes is set higher than the threshold for expenditure amount so that the number of contributors could be expected to meet the GD* threshold. In fact, for output disclosure treatment, it is more meaningful to work with specifying a reliability threshold (such as 30) in the interest of stability of resulting estimates than the traditional confidentiality threshold (10 or so) which is of course satisfied by the reliability threshold. In Section 3, we provide an objective criterion to choose suitable thresholds D* and GD* for each dataset. If the GD* threshold is reasonably high in the interest of reliability, then the problem of disclosure by dominance is considerably diffused. However, we might want to specify an additional threshold GD** to ensure that a new modified count of number of contributors with amounts more than the domain average is at or above a minimum such as 3.

Table 1B shows cell aggregates for diabetes counts and expenditure amounts that meet the thresholds D* of 50 and GD* of 10 respectively as in cell suppression for a tabular output. Cells not aggregated are safe by themselves. Cells labeled NA* correspond to suppressed cells. Choice of cell aggregation (or equivalently, choosing suppression partners such that their aggregation is safe for release) is based on the quasi-hierarchical aggregation proposed by Singh et al. (2013a). It first defines rules for category collapsing within each variable at different levels of aggregation in order depending on the need as shown in Table 2. For example, Table 2A shows four levels of age category collapsing in increasing order. After specifying rules for category collapsing within categorical variables, we need to specify the aggregation order to be implemented between variables in order to meet the threshold by cell aggregates as shown in Table 2C. Table 3 shows the aggregation summary for the example for both counts and expenditure amounts. In particular, it shows that for the full age by race by gender by diabetes table of 656 cells, the number of safe cells and cell aggregates reduces to 560 using the threshold of 50 for counts, and 396 using the threshold of 10 for modified counts. The method of quasi-hierarchical aggregation for cell suppression was motivated from the commonly used hierarchical collapsing of categories in log linear modeling for introducing various factor effects (0-dimensional or the intercept, 1-dimensional, 2-dimensional and so on). It was termed quasi to signify that in the case of cell suppression to meet the threshold, it is desired to preserve or release as many cells or cell aggregates, and therefore, category collapsing for a particular variable is not done uniformly across categories or category

combinations of other conditioning variables in a given multi-dimensional table; see Singh et al. (2013) for more details.

3. Disclosure Audit for suitable selection of thresholds D* and GD*

For the example considered in the previous section, the thresholds were chosen somewhat arbitrarily from confidentiality considerations. However, there is a need for an objective method to choose thresholds that are data-driven and provide adequate protection from disclosure or reverse-engineering. To this end, for a given threshold D* or GD* and a given high-dimensional table of counts or modified counts as the case may be, a non-parsimonious log-linear model with parameters corresponding to all linearly independent and safe cells and cell aggregates is fitted so that all the safe cells and cell aggregates are preserved at their observed true values. Once the model is fitted, all the unsafe or suppressed cells and cell aggregates can be estimated. This implies that with the usual practice of releasing a tabular output of counts with suppressed unsafe cells, an advanced user could also fit such a model and estimate all the suppressed cells.

The disclosure audit proposed by Singh et al. (2013a) entails comparing observed and estimated counts for suppressed cells and defining risk metrics based on lower quantiles (e.g., 5% or 10%) of absolute error and absolute relative error between observed and estimated counts over suppressed cells. In practice, the absolute error may be more meaningful than the absolute relative error because for small or primarily suppressed cells, the relative error may be high but the absolute error may not be adequate while for large or complementarily suppressed cells, the relative error may be small but the absolute error may be adequate. It may be noted that analogous to the classical classification error problem in gross flows, where a small classification error does not affect much the count of people who do not change their status from one time point to another but it does affect considerably the count of people who do change because the actual change is quite small, we would like to set the threshold such that the absolute relative error in complementarily suppressed cells may not be much but absolute error in primarily suppressed cells may be substantial—in practice if the 5% quantile of absolute errors is at least 1, it may be deemed adequate for primary cells. In any given application, the threshold can be revised so that a suitable risk metric is achieved for a given dataset. It is important to note that the errors in estimated suppressed cells are not released to users. It is only the data producer who can use the disclosure audit to determine a suitable threshold. Table 4B shows estimated counts for primary and secondary suppressed cells and Tables 5A and B show corresponding absolute errors and absolute relative errors. It is seen that for the threshold of 50, among the primarily suppressed cells for Hispanic females, the absolute error for most cells is at least 1 implying that the threshold of 50 seems adequate. For in-patient expenditures, the threshold of 10 for modified counts (i.e., the number of contributors) also seems adequate.

More specifically, estimates of suppressed cells are obtained by solving the following set of equations to obtain parameter estimates of a log-linear model; see Singh et al. (2013). First, we form a matrix with rows having elements of 1s and 0s with the number of elements being equal to the total number of cells in the final cross-classified table of interest; i.e., elementary cells, say M, and where rows correspond to all constraints of cells and cell aggregates. Each row has a 1 in the place that indicates which cell is included in the constraint, and 0 elsewhere. It is possible that all cells are suppressed and constraints are only in terms of margins or cell aggregates. Rows of this matrix could be linearly dependent because some constraints could be derived from others by algebraic

manipulations. However, it is sufficient to work with only linearly independent constraints. So we reduce the row dimension of the matrix to achieve linear independence. Suppose there are p independent rows and the number of columns is M—the total number of cells in the cross-classified table of interest. The estimating equations for p model parameters ($\beta's$) can be written as follows where log of the expected counts are assumed to be linear in $\beta's$, t_i denotes the ith positive constraint (value of safe cell or cell aggregate), and x_{ki} is the kth column element of the ith row of the constraint matrix taking values of 1 or 0; i=1,…, p; k= 1,…,M. There are p equations and p unknowns ($\beta's$) and the above system of nonlinear equations in principle can be solved by wellknown methods such as Newton-Raphson. However, in real applications, p can be quite large and so an alternative method of nonlinear Gauss-Seidel (Jiang, 2000) can be used. It may be noted that although all cell counts are estimated (i.e., both suppressed and non-suppressed cells) and hence all cell aggregates, cells and cell aggregates that are safe are preserved at their true values. In other words, estimated counts match safe cell and cell aggregate values (to be preserved) by construction via

$$\begin{pmatrix} x_{11}\ x_{21}\ \dots x_{k1}\ \dots x_{M1} \\ x_{12}\ x_{22}\ \dots x_{k2}\ \dots x_{M2} \\ \dots \\ \dots \\ x_{1i}\ x_{2i}\ \dots x_{ki}\ \dots x_{Mi} \\ \dots \\ \dots \\ x_{1p}\ x_{2p}\ \dots x_{kp}\ \dots x_{Mp} \end{pmatrix} \times \begin{pmatrix} \exp(x_{11}\beta_1 + x_{12}\beta_2 + \cdots x_{1i}\beta_i + \cdots x_{1p}\beta_p) \\ \exp(x_{21}\beta_1 + x_{22}\beta_2 + \cdots x_{2i}\beta_i + \cdots x_{2p}\beta_p) \\ \dots \\ \dots \\ \exp(x_{k1}\beta_1 + x_{k2}\beta_2 + \cdots x_{ki}\beta_i + \cdots x_{kp}\beta_p) \\ \dots \\ \dots \\ \exp(x_{M1}\beta_1 + x_{M2}\beta_2 + \cdots x_{Mi}\beta_i + \cdots x_{Mp}\beta_p) \end{pmatrix} = \begin{pmatrix} t_1 \\ t_2 \\ . \\ . \\ t_i \\ . \\ . \\ t_p \end{pmatrix}$$

constraints in the estimating equations. The above estimating equations coincide with maximum likelihood equations for the model parameters under the working assumption of multinomial sampling, and thus the estimates are optimal in some sense.

The above modeling is used for disclosure audit for both cell counts and modified cell counts for the expenditure variable. In the case of expenditure, although the modified cell counts are not released for fear of putting zero contributors at risk of disclosure, the estimated expenditure amounts can be released. These amounts for the suppressed cells can be estimated in a similar fashion because the expenditure variable is nonnegative; see Tables 4B and 5. Above domain estimates are examples of predictions based on aggregate or cell level modeling; see Section 8 for use of Q-PUF for general model-based predictions of domain totals where the variable need not be nonnegative.

4. Internal Consistency of Analyses Output over Repeated Queries

In any application of Q-PUF, queries may come at different times and they may refer to analysis domains formed by different combinations of variable categories with no, partial, or complete overlap of variables. In the case of same set of variables, queries may correspond to same or different analysis domains. The variables may be cross-sectional (i.e., for the same point in time), or longitudinal (i.e., for different points in time), or both. In any case we are generally dealing with a moderate number (7 ± 2 or so) of variables at a time in each query although the total number of variables could be large, and the queries are submitted sequentially. The data producer could anticipate in advance the common types of queries from which one could create an initial omnibus set of categorical variables some or all of which are expected to be part of most queries. For the

omnibus set, safe cells and cell aggregates can be determined with respect to a suitable threshold as discussed in the previous section when the outcome variable is either categorical or continuous. In the process, we can construct a checklist of variables with allowable categories for one-dimensional, two-dimensional, and higher order marginal distributions. We need two Checklists I and II: the first list for allowable domains or cell aggregates which could be cells by themselves, and the second list for non-allowable or suppressed cells. The checklist II will be useful when dealing with subsequent queries with some variables being in common with sets in earlier queries so that any suppressed complementary cell appearing in response to earlier queries continues to be suppressed. This way we can ensure internal consistency between different queries' output with respect to cell aggregation or suppression cell partners. In addition, when dealing with a sequence of queries, estimated cell counts for categorical outcome variables or cell totals for nonnegative continuous outcome variables can be ensured to be consistent with previous queries' output for common domains by enlarging the vector of safe cell or cell aggregates in equation (1) to include estimates obtained for the first time for such domains.

Besides maintaining the above internal consistency property of Q-PUF through Checklists I and II, there is another important benefit of Checklist construction and their updating over time by including more variables to satisfy user needs; this is subject to the restriction that new variables (i.e., corresponding domains defined by them) do not conflict with domains defined by existing variables in Checklist I with regard to their disclosure-safety. Thus Q-PUF is not as flexible as a general query-based analysis system for user-specified arbitrary domains but may not be limiting because Checklist I is based on common types of analysis needs. This built-in pre-screening of analytic variables provides a direct control on the types of allowable analysis, and as a result can thwart differencing attacks—a very difficult problem for query-based systems as discussed in Gomatnam et al. (2005). It follows that a user cannot game the system. For any variable or set of variables defining a rare domain submitted as a query, the system will not respond until the variables defining the domain conform to the specification in the checklist I—this is essentially another type of output consistency enforced with respect to domain definitions involving the same set of variables over different queries.

A different method termed aggregate-level PUF (Singh and Borton, 2012, and Singh et al., 2013) allows for arbitrary domain definition because it performs input disclosure treatment at an aggregate level (small groups of 10-20 individuals) although there is some loss of precision due to subsampling in comparison to Q-PUF results for domains that are preserved; i.e., considered safe for release.

5. Descriptive Inference

(a) *Inferential Estimation*: For all variables in the checklist I, it is easily seen that desired estimates for totals for domains defined by the variables will match exactly with the estimates from the original data because the domains consist of cells or cell aggregates that are preserved. This precision-preserving property of Q-PUF for estimating domain totals for variables in the checklist I carries over to more complex parameters. To see this, consider the general framework of estimating functions and the size of the support of the estimating function. In particular, for the simple problem of estimating the domain total count, let $x_{k(d)}$ denote the indicator variable for domain 'd' and $\sum_{k=1}^{N} x_{k(d)}$ the domain count parameter or $\sum_{k=1}^{N} x_{k(d)}/N$, be the domain mean or proportion parameter

to be denoted by p_d where N is the population size or the total number of beneficiaries. Then, the estimating function for estimating p_d is simply $\sum_{k=1}^{n}(x_{k(d)} - p_d)w_k$ with w_k being the sampling weight. Here the elementary estimating functions are simply$(x_{k(d)} - p_d)$. The estimating function is set to zero to obtain an estimator of the domain count as $N\sum_{k=1}^{n} x_{k(d)}w_k / \sum_{k=1}^{n} w_k$. To check for support of the estimating function, it easily follows that we need to consider the number of nonzero contributions of elementary estimating functions for a given value of the parameter p_d which is simply the number of sample observations in the domain. In the example presented earlier from DE-SynPUF, sampling weights w_k's were not considered although the estimated totals need to be weighted up to obtain population estimates. Since the original sample used for DE-SynPUF is a 5% simple random sample and a 50% simple random subsample of that was taken for the example, the resulting weights are common for all sampled units (and equal to 1/(.05)(.5)) and do not affect estimation of means. However, for unbiased estimation of population totals, they do matter in that the estimates need to be inflated by the factor 1/(.05)(.5).

Now, for the magnitude data such as the expenditure of in-patients, the estimating function for the domain total parameter is given by $\sum_{k=1}^{n}(x_{k(d)}y_k - \mu_d)w_k$ where μ_d is the domain mean expenditure over all beneficiaries. It is seen that the support of the estimating function is now given by the number of nonzero values of the product $x_{k(d)}y_k$; i.e., the number of contributors to the expenditure in the domain sample. To complete inference, variance estimate is also needed besides point estimates. If the support of the estimating function for point estimate is large enough to be safe, it is also safe for estimating variance of the estimating function, and hence for variance of the estimator.

So far we considered analysis domains deemed safe for Q-PUF. However, estimates for unsafe cells are also produced under Q-PUF as part of the disclosure audit, and if a user is interested in a domain involving unsafe cells, then the resulting estimator will not exactly match the estimator obtained from the original data. Moreover, in this case, variance estimate will need to be adjusted for estimation (or imputation) of unsafe domains. The problem is somewhat similar to the case of partial synthesis under multiple imputation, and results developed by Reiter (2003b) should be applicable. This area, however, requires further investigation.

(b) *Inferential Testing:* Comments made above for inferential estimation essentially carry over for testing purposes in an analogous manner.

6. Analytic Inference

For regression modeling at the unit or individual beneficiary level, checklist-I provides variables or collapsed versions of them that can be used for one factor, two and higher order factor effects. In other words, we only introduce covariates and corresponding regression parameters for which there is adequate estimation support from the sample. Consequently, there will be no change in inference between Q-PUF results and analysis from the original data. Despite restrictions on the analysis variables, this is not really limiting in practice from the reliability threshold point of view because allowing variables that define below threshold domain would not lead to reliable estimates anyway. Besides, the categorical collapsing needed for higher order factor effects to be included in the checklist I can be chosen in such a way that it meets requirements for common types of

analysis. There may be some restrictions that need to be imposed on releasing certain parameter estimates (not regression coefficients though) in the interest of disclosure-safety. For example, only a lower bound of the R^2-type statistics can be released to protect against disclosure situations where the modeling may be almost perfect and individual y-values can be predicted extremely well.

There is one important limitation in analysis with Q-PUF with regard to model diagnostics using residuals. The reason for this is that residuals are at the unit level, and knowledge of predicted values and the corresponding lead to disclosure of the actual value of the outcome for each individual record. To overcome this problem, an innovative method was proposed by Reiter (2003a) based on synthetic residuals after creating a synthetic microdata using multiple imputation. Alternatively, we can use the aggregate level PUF (AL-PUF., Singh et al., 2012b, and 2013b) framework of micro-groups and micro-means to compute average y-values (outcome variable) for each category of $\hat{y}$-variable (i.e., the predicted values)—here $\hat{y}$ is categorized because it is continuous. The plot of averages of the dependent (y) variable against the categorized predicted ($\hat{y}$) variable is expected to provide a reasonable approximation to the original residual plot assuming the categorization of the $\hat{y}$-variable is not too coarse.

7. Model-based Prediction of Domain Totals

After any modeling, it is natural to consider prediction of totals for domains defined by covariates or x-variables. Such predictions can be produced for any domain--allowed or not by Q-PUF. It is assumed that x-variables are available for units in the domain subpopulation for which predicted values can be obtained from the estimated model parameters and hence the predicted domain total. This estimated domain total is different from the estimated cell counts or amounts (such as expenditure) using log-linear models under descriptive inference where all the safe cell and cell aggregate values are preserved. For this estimation, the model is at an aggregate level (i.e., the elementary cell level) and not the unit level for regression modeling.

8. Concluding Remarks

The Q-PUF method for query-based analysis system formulates the problem of disclosure treatment of analysis output from queries as that of ensuring sufficient support size for estimating functions of descriptive and analytic parameters. Therefore, a version of minimum threshold criterion, termed GD*, is used for disclosure-safety. Like the usual minimum threshold criterion (D*) used for disclosure-safety of tabular data, GD* considers an underlying cross-classified table of categorical covariates used in analysis (descriptive or analytic) and the associated modified count defined as the number of contributors to the estimating function for each domain. The new criterion is adequate for aggregate level output such as domain total estimates or regression parameter estimates (hence aggregate level predictions) but for unit level output such as model residuals, extra protection is needed as a result of some transformation of the microdata itself such as synthetic residuals of Reiter (2003a) or the micro-group micro-mean representation of AL-PUF (Singh et al. 2012b).

An objective method of disclosure audit (Singh et al., 2013a) is used to decide adequacy of the threshold D* or GD*. In the process, estimates (counts of number of individual or contributor amounts for a continuous nonnegative variable) for suppressed cells are

obtained using log-linear modeling which serves to complete the table and help in a unified and simplified user interpretation by eliminating various ad hoc estimates that different users might employ.

The most difficult problem of differencing attacks in any query-based system is addressed in Q-PUF by introducing Checklists I and II for the data producer so that user queries are limited to analysis domains defined by pre-screened variables in Checklist I. Checklist I consists of all the variables defining allowable domains (i.e., safe cells and cell aggregates), and Checklist II consists of all primarily or complementary suppressed cells and cell aggregates. Checklist creation is not a one-time task. Instead it is continually updated over time starting with a comprehensive initial list based on various commonly used analysis domains anticipated by the data producer. In any subsequent output, Checklist II is used to check if there are any suppressed cells in the current output that are common with earlier output, and for the sake of internal consistency, such cells or cell aggregates are preserved at their previously estimated values by treating them as other safe cells. Moreover, no matter how many attempts are made to submit queries that do not conform to Checklist I; i.e., they have variables that in conjunction with existing variables may define unsafe domains, the system does not respond unless the query refers to variables listed in the Checklist I. This is another type of output consistency that Q-PUF maintains.

In summary, we list the four main components of Q-PUF required in any application: first, construction of Checklists I and II; second, disclosure audit to choose data-specific adequate confidentiality threshold as well as estimation of suppressed cells and cell aggregates; third, creation of an interface to communicate with the microdata via queries; and fourth, imposing additional restrictions specific to any unit level analysis output if need be.

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

Godambe, V.P. and Thompson, M.E. (1989). An extension of quasi-likelihood estimation (with discussion). Journal of Statsitical Planning and Inference, 12, 137-172.

Godambe, V.P. and Thompson, M.E. (2009). Estimating functions and survey sampling. In *Handbook of Statistics 29* (D. Pfeffermann and C.R. Rao, Eds), Vol 29B, 83-101.

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.

Jiang, J. (2000). A nonlinear Gauss-Seidel algorithm for inference about GLMM. *Computational Statistics,* 15, 229-241.

Lucero, J., Freiman, M., Singh, L., You, J., DePersio, M. and Zayatz, L. (2011). The Microdata Analysis System at the U.S. Census Bureau, SORT Special issue: Privacy in statistical databases, 2011, 77-98

McCullagh , P. and Nelder, J.A. (1989). Generalized linear models (2nd ed.) Chapman and Hall, London.

Reiter, J. P. (2003a). Model diagnostics for remote-access regression systems. *Statistics and Computing*, 13, 371-380.

Reiter, J.P. (2003b). Inference for partially synthetic public use microdata sets. *Survey methodology*, 29, 181-188.

Singh, A.C, J.M. Borton, S.H. Cohen, V.E. Welch, Jr., B. Groenhout, and Y. Lin. (2013a). Estimation for Cells Suppressed in Tabulation with Application to Output Disclosure treatment of the NSF Survey of Earned Doctorates. *ASA Surv. Res. Meth. Sec.*

Singh, A.C., Borton, J.M., Erdem, E., and Lin, Y. (2013b). Aggregate level PUF with high data confidentiality and analytic utility for general purpose analyses from Medicare claims data. *ASA Proc. Surv. Res. Meth. Sec.*

Singh, A.C, Borton, J.M., and Crego, A.M. (2012a). A generalized domain size threshold for analysis restrictions with remote analysis servers. Proceedings of the Federal Committee on Statistical Methodology, US Census Bureau. (http://www.fcsm.gov/events/papers2012.html)

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

Table 1: Untreated and Treated (or Aggregated cells) Tables of Total Counts and Expenditure by Age for: Race=Hispanic and Gender= Female (based on a 50% sample from DE-SynPUF)

(A) UNTREATED TABLES WITH SMALL CELLS

Table Variables: Race=Hispanic, Gender=Female

COUNTS		
	Diabetes	
Age	No	Yes
55	75	43
56	70	40
57	70	39
58	58	25
59	50	44
60	68	43
61	68	34
62	53	44
63	54	56
64	69	43
65	384	140
66	355	163
67	375	149
68	390	143
69	398	157
70	267	151
71	241	154
72	241	153
73	258	171
74	271	164

Shaded cells indicate need for primary

	Contributors		IP Expenditure	
	Diabetes		Diabetes	
Age	No	Yes	No	Yes
55	1	15	$17,000	$148,170
56	4	9	$60,800	$119,080
57	4	8	$46,860	$95,090
58	3	7	$18,300	$76,920
59	4	13	$72,860	$265,500
60	1	14	$10,000	$271,340
61	2	9	$33,000	$188,820
62	2	14	$13,000	$247,500
63	4	17	$49,000	$336,310
64	3	9	$16,000	$164,220
65	9	21	$119,180	$197,240
66	5	41	$73,850	$478,890
67	12	29	$207,980	$439,390
68	12	27	$105,830	$339,530
69	8	25	$174,500	$443,660
70	8	32	$109,900	$396,890
71	6	38	$71,000	$730,100
72	12	27	$96,310	$380,910
73	10	35	$247,000	$572,580
74	6	31	$135,000	$422,730

Shaded cells indicate need for primary suppression, <10 non-zero contributors to the total.

(B) AGGREGATED TABLES

Table Variables: Race=Hispanic, Gender=Female

COUNTS		
	Diabetes	
Age	No	Yes
55		
56		
57	323	191
58		
59		
60		
61		
62	312	220
63		
64		
65	384	140
66	355	163
67	375	149
68	390	143
69	398	157
70	267	151
71	241	154
72	241	153
73	258	171
74	271	164

	Contributors			
	Diabetes		Diabetes	
Age	No	Yes	No	Yes
55				
56				
57	16	52	$215,820	$704,760
58				
59				
60				
61				
62	12	63	$121,000	$1,208,190
63				
64				
65				
66				
67	53	254	$681,340	$1,898,710
68				
69				
70				
71				
72	36	161	$659,210	$2,503,210
73				
74				

*IP expenditure is not availabe at this level of race crossed with age, gender and diabetes. It is aggregated with other levels of race

NOTE: Tables 1, 4 and 5 have been abbreviated to a subset of ages in the interest of space.

Table 2: Category Collapsing within a Variable and Aggregation Order between Variables
(Variables gender and diabetes have only two categories and are not collapsed0

2A: Age Variable

Variable	1st Level (41 Categories)	2nd Level (21)	3rd level (9)	4th Level (4)
Age	50 and Under	54 and under	54 and under	64 and under
	51-89 by single year	50-59, 60-64	50-89 by 5 year intervals	65-74
	90 and over	65-79 by single year	90 and over	75-84
		80-84, 85-89		85 and over
		90 and over		

2B: Race Variable

Variable	1st Level (4)	2nd Level (3)	3rd level (2)
Race	White	White	White
	Black	Black	Other
	Hispanic	Other	
	Other		

2C: Cell Aggregation order between Variables

Step	Variable	Level Move
1	Age	1 to 2
2	Age	2 to 3
3	Race	1 to 2
4	Age	3 to 4
5	Race	2 to 3

Table 3: Aggregation Summary for the DE-SynPUF Example

1 Dimension		
	Count (min=50)	IP Expenditure Amount (min=10)
age (41 categories)) race (4 categories) gender (2) diabetes (2)	no aggregation needed	no agg needed
2 Dimensions		
age x race (164) age x gender (82) age x diabetes (82) race x gender (8) race x diabetes (8) gender x diabetes (4)	no agg needed	no agg needed
3 Dimensions		
age x race x gender (328) age x race x diabetes (328) age x gender x diabetes (164) race x gender x diabetes (16)	no agg needed	age (lev2) (312) age (lev 2) (264) no agg needed no agg needed
4 Dimensions		
age x race x gender x diabetes (656)	age (lev 2) (560)	age (lev 3), race (lev 2) (396)

Note: When there is aggregation it means that some categories of the aggregated variable are combined in order to make a combination that meets the threshold requirement. In this example, the minimum for diabetes counts is 50 while the minimum for contributors to the mean or total of the continuous expenditure variable is 10. Detailed information about the number of cell aggregations for each combination of variables to give an idea of how much aggregation is taking place Is not shown to save space

Table 4: Observed and Estimated Diabetes Counts and Expenditure Amounts

(A) UNTREATED TABLES WITH SMALL CELLS

Table Variables: Race=Hispanic, Gender=Female

COUNTS		
	Diabetes	
Age	No	Yes
55	75	43
56	70	40
57	70	39
58	58	25
59	50	44
60	68	43
61	68	34
62	53	44
63	54	56
64	69	43
65	384	140
66	355	163
67	375	149
68	390	143
69	398	157
70	267	151
71	241	154
72	241	153
73	258	171
74	271	164

Shaded cells indicate need for primary

IP Expenditure		
	Diabetes	
Age	No	Yes
55	$17,000	$148,170
56	$60,800	$119,080
57	$46,860	$95,090
58	$18,300	$76,920
59	$72,860	$265,500
60	$10,000	$271,340
61	$33,000	$188,820
62	$13,000	$247,500
63	$49,000	$336,310
64	$16,000	$164,220
65	$119,180	$197,240
66	$73,850	$478,890
67	$207,980	$439,390
68	$105,830	$339,530
69	$174,500	$443,660
70	$109,900	$396,890
71	$71,000	$730,100
72	$96,310	$380,910
73	$247,000	$572,580
74	$135,000	$422,730

Shaded cells indicate need for primary suppression, <10 non-zero contributors to the total.

(B) ESTIMATED TABLES

Table Variables: Race=Hispanic, Gender=Female

COUNTS		
	Diabetes	
Age	No	Yes
55	*76.9*	*41.1*
56	*67*	*43*
57	*67.6*	*41.4*
58	*55.6*	*27.4*
59	*56*	*38*
60	*65*	*46*
61	*64.5*	*37.5*
62	*54*	*43*
63	*57.7*	*52.3*
64	*70.8*	*41.2*
65	384	140
66	355	163
67	375	149
68	390	143
69	398	157
70	267	151
71	241	154
72	241	153
73	258	171
74	271	164

IP Expenditure		
	Diabetes	
Age	No	Yes
55	*$37,196*	*$127,974*
56	*$44,136*	*$135,744*
57	*$38,164*	*$103,786*
58	*$13,703*	*$81,517*
59	*$82,620*	*$255,740*
60	*$27,142*	*$254,198*
61	*$14,801*	*$207,019*
62	*$32,798*	*$227,702*
63	*$33,422*	*$351,888*
64	*$12,837*	*$167,383*
65	*$82,006*	*$234,414*
66	*$103,118*	*$449,622*
67	*$201,113*	*$446,257*
68	*$126,673*	*$318,687*
69	*$168,430*	*$449,730*
70	*$114,960*	*$391,830*
71	*$108,831*	*$692,269*
72	*$104,489*	*$372,731*
73	*$200,805*	*$618,775*
74	*$130,126*	*$427,604*

NOTE: Tables 1, 4 and 5 have been abbreviated to a subset of ages in the interest of space.

Table 5: Estimation Errors in Diabetes Counts and Expenditure Amounts

(A) ABSOLUTE ERROR BETWEEN ESTIMATED AND REAL VALUES

Table Variables: Race=Hispanic, Gender=Female

COUNTS		
	Diabetes	
Age	No	Yes
55	1.86	1.86
56	3.04	3.04
57	2.42	2.42
58	2.35	2.35
59	5.96	5.96
60	2.99	2.99
61	3.54	3.54
62	1.01	1.01
63	3.73	3.73
64	1.79	1.79
65		
66		
67		
68		
69		
70		
71		
72		
73		
74		

Shaded cells indicate need for primary suppression, count<50.

IP Expenditure		
	Diabetes	
Age	No	Yes
55	$ 20,196	$ 20,196
56	$ 16,664	$ 16,664
57	$ 8,696	$ 8,696
58	$ 4,597	$ 4,597
59	$ 9,760	$ 9,760
60	$ 17,142	$ 17,142
61	$ 18,199	$ 18,199
62	$ 19,798	$ 19,798
63	$ 15,578	$ 15,578
64	$ 3,163	$ 3,163
65	$ 37,174	$ 37,174
66	$ 29,268	$ 29,268
67	$ 6,867	$ 6,867
68	$ 20,843	$ 20,843
69	$ 6,070	$ 6,070
70	$ 5,060	$ 5,060
71	$ 37,831	$ 37,831
72	$ 8,179	$ 8,179
73	$ 46,195	$ 46,195
74	$ 4,874	$ 4,874

Shaded cells indicate need for primary suppression, <10 non-zero contributors to the total.

(B) RELATIVE ERROR BETWEEN ESTIMATED AND REAL VALUES

Table Variables: Race=Hispanic, Gender=Female

COUNTS		
	Diabetes	
Age	No	Yes
55	2.4%	4.5%
56	4.5%	7.1%
57	3.6%	5.8%
58	4.2%	8.6%
59	10.6%	15.7%
60	4.6%	6.5%
61	5.5%	9.4%
62	1.9%	2.4%
63	6.5%	7.1%
64	2.5%	4.3%
65		
66		
67		
68		
69		
70		
71		
72		
73		
74		

IP Expenditure		
	Diabetes	
Age	No	Yes
55	54%	16%
56	38%	12%
57	23%	8%
58	34%	6%
59	12%	4%
60	63%	7%
61	123%	9%
62	60%	9%
63	47%	4%
64	25%	2%
65	45%	16%
66	28%	7%
67	3%	2%
68	16%	7%
69	4%	1%
70	4%	1%
71	35%	5%
72	8%	2%
73	23%	7%
74	4%	1%

NOTE: Tables 1, 4 and 5 have been abbreviated to a subset of ages in the interest of space.