Maintaining Analytic Quality while Protecting Confidentiality of Survey Data

Avi Singh

Carleton University, Ottawa

NISS/NCHS Data Confidentiality Workshop,

May 1-2 ’08, Hyattsville, MD

Outline

- Disclosure-treated PUFs and their structure
- Inside vs. Outside Intrusion Scenarios
- Disclosure Risk from Inside Intrusion
- Simultaneous Control on Disclosure Risk and Information Loss
- PUF for Survey Weighted Data

Outline

- Probabilistic Measures of Disclosure Risk and Information Loss
- Survey Sampling Based Disclosure Treatment via MASSC
- Illustrative Example on MASSC and Associated Disclosure Risk and Analytical Quality Measures
- Analysis of MASSC-treated Data Set
- Comparison with Alternative Methods and Concluding Remarks

Disclosure Treated Public Use Files (PUFs)

- Great user demand of administrative, census, and sample survey data; typically collected under confidentiality pledge.
- Highly sensitive data not previously available to researchers can be made available after disclosure treatment.
- In data mining applications for detecting rare events or characteristics of small subgroups, disclosure treated surrogate data can be used by researchers at large, and then the final analysis on the original dataset can be performed under tight security.

Structure of PUF

Analytic Variables (AVs):

- **(Indirect disclosure) Identifying Variables (IVs):** provide, in general, information about demographic, geographic, and socio-economic status, e.g.,
  - for BRFSS (behavior risk factor surveillance survey) data, age group, race, gender, education, income, height, weight, freq of eating fruits, flu shots, etc.

Structure of PUF

Analytic Variables (AVs):

- **(Disclosure) Sensitive Variables (SVs):** provide, in general, information about medical, financial, social, and professional status, e.g.,
  - for BRFSS data, asthma condition, diabetes condition, # permanent teeth removed, drinking alcohol and driving car, reason for HIV test, method of birth control, etc.
- **(Analytical) Quality Control Variables (QCVs):** correspond to study variables such as asthma condition for an age group.
- Need groups of key IVs and SVs for measuring disclosure risk, and key QCVs for measuring information loss.

Inside Intrusion Scenario

- Target’s presence in the database is known to the intruder.
- “Disclosure by response knowledge” – an important inside intrusion scenario from respondent’s perspective; Bethlehem et al. (1990)
- Respondent identifies his own record and is concerned about its disclosure by someone who might know enough about him to identify his record.
- Reputation and credibility of a data producer are at stake if the respondents in the database do not have confidence in the producer.

Inside Intrusion Scenario ...Ctd.

- Coalition intrusion in tabular data may put at risk an individual belonging to cells with small counts.

- Child’s response to drug behavior in NSDUH survey data at risk of disclosure to parents.

- For the administrative data (e.g., Canadian Cancer Registry), neighbor or a coworker might know about the cancer episode.

Inside Intrusion Scenario ...Ctd.

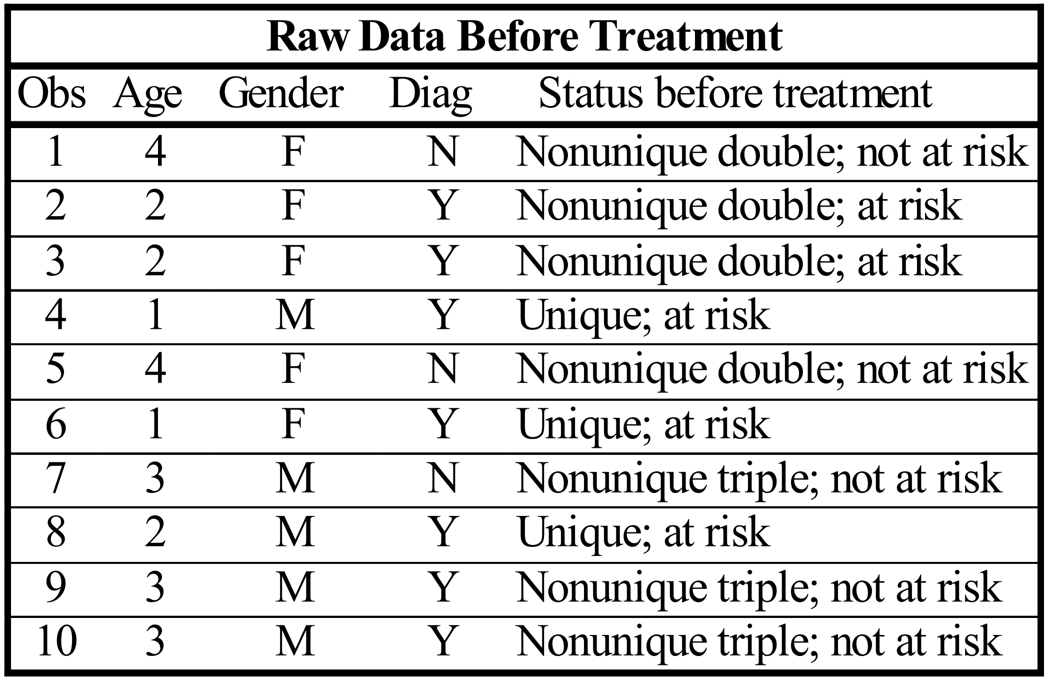
- Risk is 100% for unique records with sensitive values, as well as for nonunique records with common sensitive values of a SV.
- Therefore, some disclosure treatment via perturbation (substitution of IVs) and suppression (subsampling-out part or whole records) is necessary.

Outside Intrusion Scenario

- Target’s presence in the database unknown to the intruder.
- Here risk for a sample unique is not 0/1; can be estimated under a model for population uniques, Skinner and Holmes (1998).
- Disclosure treatment may not be necessary.
- Inside intrusion scenario puts more onus on the data producer.
- Protecting against inside intrusion automatically protects against outside intrusion; provides an upper bound on disclosure risk.

Disclosure Risk from an Inside Intruder: an example

- Consider a hypothetical data with 10 observations consisting of two risk strata of uniques and nonuniques.
  (IVs= age, gender; SV=Asthma condition)

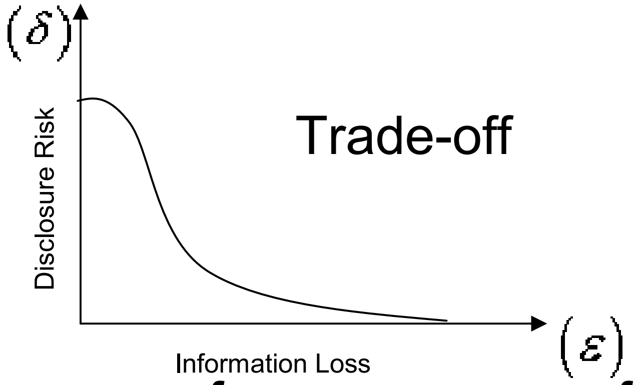


Raw Data Before Treatment

Obs	Age	Gender	Diag	Status before treatment
1	4	F	N	Nonunique double; not at risk
2	2	F	Y	Nonunique double; at risk
3	2	F	Y	Nonunique double; at risk
4	1	M	Y	Unique; at risk
5	4	F	N	Nonunique double; not at risk
6	1	F	Y	Unique; at risk
7	3	M	N	Nonunique triple; not at risk
8	2	M	Y	Unique; at risk
9	3	M	Y	Nonunique triple; not at risk
10	3	M	Y	Nonunique triple; not at risk

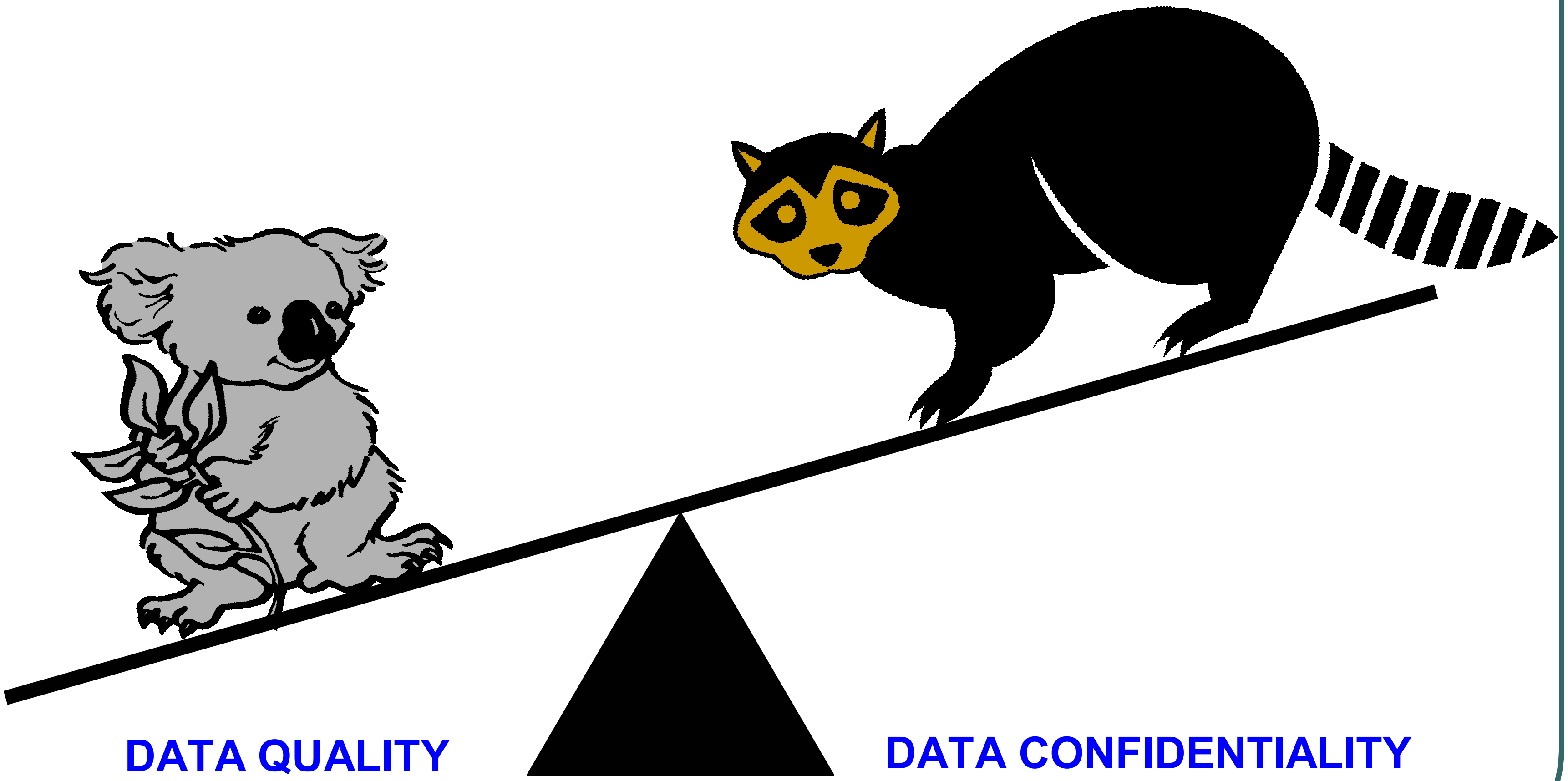
Simultaneous Control on Disclosure Risk and Information Loss: A Conundrum

- Any disclosure treatment leads to information loss.
- How to balance the tension between disclosure risk (δ) due to limited amount of perturbation and suppression, and information loss (ε) due to introduction of bias and variance?



- Useful to have (ε,δ) measures for any process of disclosure treatment; similar to Risk-Utility framework of Duncan and Lambert (1986), and Trottini and Fienberg (2002); also Skinner and Carter (2003), Winkler (2004).

Disclosure: A Balancing Act



DATA QUALITY

DATA CONFIDENTIALITY

PUF for Survey Weighted Data

- Stratum/PSU subset identification may act as IVs but needed for variance estimation. However, sufficient to have pseudo identifiers. Then, there may not be any disclosure issue if units within the subset are treated.

- Sampling weight for each unit is important for reducing selection bias in analysis (Pfeffermann, 1993), but it may also act as an IV due to over/under sampling. However, after calibration for nonresponse, coverage bias/post-stratification, and extreme values, as well as possibly a second stage of calibration post-disclosure treatment, its value as an IV may be negligible.

- See De Waal and Willenborg (1997) for potential problems in general.

Disclosure Treatment: Deterministic vs. Stochastic

Deterministic Selection for Treatment: All records at risk are treated; the risk goes to zero but it may lead to high information loss. Also, there is no protection against new IVs.

Stochastic Selection for Treatment: All records are subject to treatment but only a small random subset is actually treated; leads to low information loss and protection against new IVs. However, risk is not zero but small after treatment.

Note: Need a probabilistic/stochastic framework to measure and control disclosure risk and information loss.

Probabilistic Measures of Disclosure Risk

Suppose the data set is divided into risk strata comprising uniques, and nonuniques

π_h: proportion of records in stratum h

ψ_h : substitution rate in stratum h

ϕ_h ; subsampling rate in stratum h

χ_h : probability of misclassification of a record from stratum h

ξ_u : probability that value of SV is not sensitive given that it appears unique and survived treatment.

ξ_{nu} : probability that given that the record appears nonunique and survived treatment, no SV has a common value or SV may have common SV value but insensitive.

Disclosure Risk Measure

- $\pi_U, \pi_{NU}, \psi_U, \psi_{NU}$ are known before disclosure treatment. However, ϕ_U, ϕ_{NU} are not known in advance of the treatment but can be estimated from the realized sample after substitution treatment. They can be estimated more precisely by Monte Carlo simulation of substitution. No strong modeling assumptions are needed.

•The probabilities $\chi_U, \chi_{NU}, \xi_U, \xi_{NU}$ are also not known and can be estimated by simulations of substitution and subsampling.

• For obtaining stable estimates of above probabilities with a single treated dataset, assume that they are common over profiles or subgroups of records.

Disclosure Risk Measure

- Risk measure for a unique looking record in the treated dataset:

$$\delta_u = \pi_U(1-\psi_U)\phi_U(1-\chi_U)(1-\zeta_u) \\ +\pi_{NU}(1-\psi_{NU})\phi_{NU}\chi_{NU}(1-\zeta_u)$$

- For a nonunique looking record:

$$\delta_{nu} = \pi_{NU}(1-\psi_{NU})\phi_{NU}(1-\chi_{NU})(1-\zeta_{nu}) \\ +\pi_U(1-\psi_U)\phi_U)\chi_U(1-\zeta_{nu})$$

- Overall risk: $\delta = \max\{\delta_u, \delta_{nu}\}$

Information Loss Measure

- By regarding the original data set as the finite population, and the treated one as a sample, we can compute RB and RRMSE of several study variables (z) and compute the overall information loss as

$$\varepsilon = \max_z \{RRMSE(\hat{\theta}_z)\}$$

- Mean Square Error:

$$\begin{aligned} MSE(\hat{\theta}_{z^*}) &= E_{\psi\phi}\left(\hat{\theta}_{z^*} - \theta_z\right)^2 \\ &= E_\psi V_{\phi|\psi}(\hat{\theta}_{z^*}) + E_\psi B^2_{\phi|\psi}(\hat{\theta}_{z^*}) \\ &\leq \left(\beta + \alpha\right)\theta_z^2 = \varepsilon^2\ \theta_z^2 \end{aligned}$$

- Obtain other analytical quality measures by computing relative differences of estimates of finite population and superpopulation (model) parameters before and after treatment.

A Survey sampling-based Method for Disclosure treatment

- Under inside intrusion, database is the finite population.

- Subtle analogy between census taking (with associated high monetary cost) and releasing the original database (with associated high disclosure cost).

A Survey sampling-based Method for Disclosure treatment

- Taking a well-designed sample from the finite population vs. treating a database for disclosure:

  - Stratify for over/under sampling vs. create risk strata for over/under treatment)
  - Impute for item nonresponse vs. perturb at random
  - Sample selection vs. random non-suppression
  - Weight calibration to reduce bias due to nonresponse and variance due to sampling common for both scenarios.

- Re: Singh (2002, 2006), Singh, Yu, and Dunteman (2003), Singh, Yu, and Wilson (2004)

Process of MASSC
(A nonsynthetic approach)

Steps:

I: Micro Agglomeration
(for creating risk strata to check & control the number of records at risk of disclosure after categorizing IVs if necessary.)

II: Optimal Random Substitution
(to introduce uncertainty primarily about the identity of a target)

III: Optimal Random Subsampling
(to introduce uncertainty primarily about the presence of a target.)

IV: Optimal Calibration
(to reduce bias due to substitution and variance due to subsampling.)

Note: Can't have PUF with finer categories of IVs than used in the disclosure treatment.

A Simple Illustrative Example (Micro Agglomeration)

Raw Data			
Obs	Age	Gender	Diag
1	4	F	N
2	2	F	Y
3	2	F	Y
4	1	M	Y
5	4	F	N
6	1	F	Y
7	3	M	N
8	2	M	Y
9	3	M	Y
10	3	M	Y

Data After Micro Agglomeration				
Obs	Age	Gender	Diag	Status before treatment
4	1	M	Y	Unique; at risk
6	1	F	Y	Unique; at risk
8	2	M	Y	Unique; at risk
2	2	F	Y	Nonunique double; at risk
3	2	F	Y	Nonunique double; at risk
1	4	F	N	Nonunique double; not at risk
5	4	F	N	Nonunique double; not at risk
9	3	M	Y	Nonunique triple; not at risk
7	3	M	N	Nonunique triple; not at risk
10	3	M	Y	Nonunique triple; not at risk

Note: Under Inside Intrusion, unique records with sensitive values are at risk, and nonunique records with common sensitive values of a SV are at risk.

A Simple Illustrative Example (Substitution)

Data After Micro Agglomeration					After Substitution			
Obs	Age	Gender	Diag	Status before treatment	Age	Gender	Diag	
4	1	M	Y	Unique; at risk	1	M	Y	
6	1	F	Y	Unique; at risk	1	M	Y	
8	2	M	Y	Unique; at risk	2	M	Y	
2	2	F	Y	Nonunique double; at risk	2	F	Y	
3	2	F	Y	Nonunique double; at risk	2	F	Y	
1	4	F	N	Nonunique double; not at risk	4	F	N	
5	4	F	N	Nonunique double; not at risk	3	M	N	
9	3	M	Y	Nonunique triple; not at risk	3	M	Y	
7	3	M	N	Nonunique triple; not at risk	3	M	N	
10	3	M	Y	Nonunique triple;not at risk	2	M	Y	

A Simple Illustrative Example (Subsampling)

Data After Micro Agglomeration					After Substitution			After Subsampling
Obs	Age	Gender	Diag	Status before treatment	Age	Gender	Diag	Status after treatment
4	1	M	Y	Unique; at risk	1	M	Y	Sampled out
6	1	F	Y	Unique; at risk	1	M	Y	Pseudo-unique
8	2	M	Y	Unique; at risk	2	M	Y	Pseudo-nonunique double
2	2	F	Y	Nonunique double; at risk	2	F	Y	Pseudo-unique
3	2	F	Y	Nonunique double; at risk	2	F	Y	Sampled out
1	4	F	N	Nonunique double; not at risk	4	F	N	Pseudo-unique
5	4	F	N	Nonunique double; not at risk	3	M	N	Pseudo-nonunique triple
9	3	M	Y	Nonunique triple; not at risk	3	M	Y	Nonunique triple
7	3	M	N	Nonunique triple; not at risk	3	M	N	Nonunique triple
10	3	M	Y	Nonunique triple;not at risk	2	M	Y	Pseudo-nonunique double

A Simple Illustrative Example (Calibration)

Data After Micro Agglomeration					After Substitution			After Subsampling	After Calibration
Obs	Age	Gender	Diag	Wt	Age	Gender	Diag	Status after treatment	Wt
4	1	M	Y	1	1	M	Y	Sampled out	0.00
6	1	F	Y	1	1	M	Y	Pseudo-unique	0.83
8	2	M	Y	1	2	M	Y	Pseudo-nonunique double	0.83
2	2	F	Y	1	2	F	Y	Pseudo-unique	2.50
3	2	F	Y	1	2	F	Y	Sampled out	0.00
1	4	F	N	1	4	F	N	Pseudo-unique	2.50
5	4	F	N	1	3	M	N	Pseudo-nonunique triple	0.83
9	3	M	Y	1	3	M	Y	Pseudo-nonunique triple	0.83
7	3	M	N	1	3	M	N	Pseudo-nonunique triple	0.83
10	3	M	Y	1	2	M	Y	Pseudo-nonunique double	0.83

A Simple Illustrative Example (MASSC Result)

Raw Data			
Obs	Age	Gender	Diag
1	4	F	N
2	2	F	Y
3	2	F	Y
4	1	M	Y
5	4	F	N
6	1	F	Y
7	3	M	N
8	2	M	Y
9	3	M	Y
10	3	M	Y

Data After MASSC				
Obs	Age	Gender	Diag	Wt
6	1	M	Y	0.83
8	2	M	Y	0.83
2	2	F	Y	2.50
1	4	F	N	2.50
5	3	M	N	0.83
9	3	M	Y	0.83
7	3	M	N	0.83
10	2	M	Y	0.83

Measure of Disclosure Risk (δ)

Record Appears	Risk Strata	π	ψ	ϕ	χ	ζ
Unique	U	3/10	1/3	2/3	1/1	1/2
	NU	7/10	2/7	6/7	2/4	
Nonunique	U	3/10	1/3	2/3	1/1	2/3
	NU	7/10	2/7	6/7	2/4	

$$\hat{\delta}_u = \pi_U(1-\psi_U)\hat{\phi}_U(1-\hat{\chi}_U)(1-\hat{\zeta}_u) + \pi_{NU}(1-\psi_{NU})\hat{\phi}_{NU}\hat{\chi}_{NU}(1-\hat{\zeta}_u) \quad = 0 + 0.1071$$

$$\hat{\delta}_{nu} = \pi_{NU}(1-\psi_{NU})\hat{\phi}_{NU}(1-\hat{\chi}_{NU})(1-\hat{\zeta}_{nu}) + \pi_U(1-\psi_U)\hat{\phi}_U\hat{\chi}_U(1-\hat{\zeta}_{nu})$$

$$= (\frac{7}{10})(1-\frac{2}{7})(\frac{6}{7})(1-\frac{2}{4})(1-\frac{2}{3}) + (\frac{3}{10})(1-\frac{1}{3})(\frac{2}{3})(\frac{1}{1})(1-\frac{2}{3}) \quad = 0.0714 + 0.0444 = 0.1158$$

Measure of Information Loss (ε)

	obs	v	ψ	z*	ϕ
U	4	0	1/3	1	2/3
	6	1		1	
	8	0		1	
NU	2	0	2/7	0	6/7
	3	0		0	
	1	0		0	
	5	0		0	
	9	0		1	
	7	0		0	
	10	0		1	

- Study variable (z): Asthma among males, $N^{-1}\theta_z = 0.40$
- Expected Bias squared

  $= N^{-2}[\sum_h N_h(1-\psi_h)\psi_h S^2_{v,h} + (\sum_h(\sum_i v_i)\psi_h)^2]$

  = 1/300, where $v_i = \tilde{z}_i - z_i$
- Variance

  $= N^{-2}\sum_h N_h(1/\phi_h - 1)S^2_{z^*,h}$

  = 1/360
- RRMSE = 0.195 $\varepsilon = \max_z\{RRMSE(\hat{\theta}_z)\}$

Analysis with MASSC-treated Data

- Descriptive parameters (inference about FPQs for the parent population of the original sample data set)
  - Interest in PE, VE, and IE for means and totals
  - Account for two phase sampling. Can use single phase methods by making selection of records for substitution and subsampling conditionally independent across (pseudo) PSUs given the first phase sample.
  - Variance estimation adjusted for substitution (or imputation) of IVs; nonstandard because imputation flags not part of PUF, also account for multivariate relationships..

Analysis with MASSC-treated Data

- Model parameters (inference about superpopulation)
  - Define EFs (estimating functions) which are like FPQs(finite population quantities), and then proceed as above.
  - Both IVs and SVs are categorized for MASSC treatment; models for discrete data applicable, may be adequate in practice.

Alternative Methods

- Tabular data of counts
  - Specific structure as SVs are part of the cross-classification based on IVs. Scope of alternative treatment based on controlled perturbation of cell counts; see e.g., Cox, Kelly, and Patil (2004).
- Synthetic data
  - Based on Rubin’s multiple imputation idea; theoretically appealing ; e.g., Raghunathan, Reiter, and Rubin (2003), Little and Liu (2002).
  - However, difficult to create high dimensional models for survey data with informative designs; Reiter, Raghunathan, and Kinney (2006).
  - Multiple imputation after inverse sampling to undo the complex data structure seems promising for synthetic data; Hinkins, Oh, Scheuren (1997), Rao, Scott, and Benhin (2003).

Concluding Remarks

- MASSC being a survey sampling based method is applicable to any data (macro or micro) as long as it can be represented as a record level file that can be sampled. It gives rise to nonsynthetic disclosure treated data.

- Applicable to longitudinal data; may need to revise substitution and subsampling rates in view of additional IVs.

- With data at different levels (individual, family, and household), use MASSC treatment for each level.

Concluding Remarks

- Some areas of concern are categorization of IVs in MASSC-treated data, and suitable adjustment of data analysis for imputation or substitution.

- Direction of future research: Use of synthetic methods to produce multiple copies of substituted data set before subsampling might overcome some concerns. Here substitution IVs is done only for those records randomly chosen under given selection rates. Under the inside intrusion scenario, the original data set is regarded as the population and so the sampling design can be ignored for modeling required under MI.

Concluding Remarks

- However, for analysis, sampling weights after combining first and second phases need to be taken into account.
- Although IVs are categorized for finding records at risk among uniques and nonuniques in order to define risk strata and corresponding treatment rates, the noncategorized IVs can be released in PUF if synthetic substitution is used under MASSC.
- There is potential for unifying strengths of synthetic and nonsynthetic approaches.

Thank you