An Evaluation of the Impact of Missing Data on Disclosure Risk Measures

Tom Krenzke[1], Jianzhu Li[1], Lin Li[1]
[1]Westat, 1600 Research Blvd, Rockville, MD 20815

Abstract

This paper focuses on measuring disclosure risk when missing data exists among key identifying variables. It is well known that combining identifying variables together can lead to the identification of an individual. Records that are unique in the sample based on a set of identifiers may not be ‘true’ uniques if there exists at least one other record that is a match on a non-missing subset of variables, because it is unknown if the true values match among the missing subset of variables. Therefore, there is some protection from missing values due to the uncertainty about their true values, and it is unclear how much protection is provided by the missing data items. In addition, available software handles missing data differently when measuring disclosure risk. In this paper we describe an approach to help gauge the impact of missing data on disclosure risk measures. We conduct an illustration of risk measures using public use data, and conduct a simulation to evaluate the missing value impact.

Key Words: Confidentiality, sample unique, statistical disclosure control

1. Introduction

We present a data dissemination scenario that leads to a public use file (PUF) and where the disclosure risk is to be measured. The need to limit data disclosure risk prior to releasing a PUF is driven partly by laws, including the Privacy Act (1974), which protects records maintained on individuals. Further motivation is fear -- fear that releasing data will run the risk of breaches, where trust and response rates to surveys may plummet, or harm could result to individuals. Several risk scenarios that provide examples of data intruders are in El Emam, et al. (2009). For example, the authors discuss a prosecutor scenario, where the data intruder is looking for a specific person, and a journalist scenario, where the intruder is not looking for a specific person, but is motivated to break a story about a breach. In general, disclosure risk may arise if a data intruder intends to identify individuals and disclose their identities or attributes through the matching of known attributes. We will begin by discussing some risk measures and show results from a multi-country survey. Then with assumptions made in a simulation, we show how the presence of missing data can cause a significant overestimate of the disclosure risk, and provide a solution to improve the estimate.

2. Risk Measures

For all disclosure risk approaches, combinations of variables need to be reviewed since only a few variables are needed to define a unique record in the sample, which is referred to as a sample unique. Identifying sample uniques, or sparse combinations of variables, and gauging their uniqueness in the population is a goal of the disclosure risk measures discussed in this paper. Successful identification of high risk values in the data leads to more efficient disclosure control by targeting the data values associated with high risk. The following are summaries of measures considered in this paper, which are described further in Li and Krenzke (2013).

The Exhaustive Tabulations Assessment (ETA) identifies sample uniques or sparse combinations of variables through exhaustive tabulations. ETA was created and is used at Westat and some government agencies for identifying high risk values. The key variables used for forming the tables are usually indirect identifiers. The ETA algorithm scans exhaustively through all possible multi-dimensional tables defined from the list of key variables, and identifies the records that fall into the sparse cells. Sparse cells refer to the table cells with a cell count (weighted or unweighted) being less than a given threshold rule. If the threshold rule is 2 (unweighted), then the ETA algorithm identifies the sample uniques. The ETA attempts to relate risk in the sample tabulations to the population through setting thresholds based on the sampling weight. A risk score is computed for each data record as the number of times that a record contributes to the sparse cells (number of violations) after scanning and identification are completed.

The Special Unique Detector Algorithm (SUDA) conducts a more intelligent search than the ETA. Elliot, et al. (1998) observed that special uniques have higher chances of being population uniques than sample uniques that are not special (i.e., random uniques). Special uniques are sample uniques on coarser, less detailed attributes, and therefore are more risky than the sample uniques on finer, more detailed attributes, based on the premise that less information to identify a sample unique is more risky. The unique attribute set of a special unique is called minimal sample unique (MSU). The MSU is a set of attributes that can uniquely identify a data record, but none of its subsets are unique. On the other hand, every superset of a unique attribute set, MSU or otherwise, must also be unique. This is referred to as Superset Relationship by Elliot, et al. (2002). The risk of a given record being population unique increases as the size of the MSUs decreases and the number of the MSUs increases. There are different versions of SUDA scores, and we describe and use the following version:

$$\text{SUDA}=\sum_{s=1}^{M}\left(R_s \prod_{j=s}^{M}(ATT-j)\right)$$

where, M is the user-specified maximum size of attribute set, ATT is the total number of attributes in the dataset, R_s is the number of MSUs of size s. The number of MSUs of size s is weighted by the number of distinct paths from the current attribute set to the superset of the user-specified maximum size M.

We also studied two approaches to estimate re-identification risk. Re-identification disclosure occurs when an intruder correctly matches a target individual in a sample and a unit in the population by an available list of key variables, and identifies the individual. There are several approaches to measuring re-identification risk, including probability-based linkage that involves matching the sample file to a population file. An algorithm is discussed in Jaro (1989), and summaries are found in Winkler (1993) and Domingo-Ferrer and Torra (2001). However, the data disseminators may not have access to the population file, and therefore risk measures were developed that explicitly acknowledge the sample was selected from a population, and makes use of the sampling weight and models to estimate the re-identification risk. In a multiple dimensional table defined by a set of key variables, let F_k and f_k be the population count and the sample count in cell k, $k = 1, \ldots, K$, respectively. Under the above assumptions, the probability of re-identification of individual i being in cell k takes the form $1/F_k$ when F_k is known to the intruder (Duncan and Lambert, 1989). The risk is maximum if $F_k = f_k = 1$. Benedetti and Franconi (1998) proposed that the uncertainty on F_k is accounted for by introducing the distribution of the population counts given the sample counts. In this paper, we consider re-identification risk as the probability that a sample unique is a population unique.

Re-identification risk is computed in the Mu-Argus software (available at http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf, accessed April 14, 2014), which was developed in Europe in a collaborative effort involving Statistics Netherlands. The risk formulas originally come from the negative binomial distribution from Benedetti and Franconi (1998), then Polletini (2003) derived formulas based on approximations of the hypergeometric function. The approximate risk for a cell count equal to 1, or $f_k = 1$, is $r_i = -log(\hat{\pi}_k)\frac{\hat{\pi}_k}{1-\hat{\pi}_k}$. The approximations are extended to cell counts greater than 1.

Skinner and Shlomo (2008) developed and investigated approaches to specifying log-linear models that can be used in practice for risk assessment. Focusing on sample unique cases, a risk measure can be expressed as $r_{1i} = P(F_k = 1 | f_k = 1)$, where, the ith record in the data falls into the kth cell of the table created by the set of key variables. Considering similar model assumptions as previous work (Bethlehem, et al. 1990), suppose F_k follows independently Poisson model of parameter λ_k, $F_k \sim \mathrm{P}(\lambda_k)$. The sample is drawn by Bernoulli sampling with known inclusion probability π_k, $f_k | F_k \sim \mathrm{Binom}[F_k, \pi_k]$. Then the sample counts f_k also independently follow the Poisson distribution, $f_k \sim \mathrm{P}(\pi_k \lambda_k)$. According to the Bayes Theorem, $F_k | f_k \sim \mathrm{P}[\lambda_k(1 - \pi_k)] + f_k$. The risk measure then becomes $r_{1i} = exp(-\lambda_k(1 - \pi_k))$. Further assume λ_k are related via the log-linear model, which allows “borrowing strength” between table cells: $\log \lambda_k = \boldsymbol{x}'_k \boldsymbol{\beta}$, where, x_k is a vector depending upon the values of key variables in cell k, and $\boldsymbol{\beta}$ is the parameter vector. Typically x_k include main effects and low-order interaction terms of the key variables. The risk measures above can be estimated by replacing λ_k by $\hat{\lambda}_k = exp(\boldsymbol{x}'_k \hat{\boldsymbol{\beta}})$.

3. Risk Measure Illustration Using PUF Data

To illustrate the impact on risk levels of including or not including certain variables in the PUF among the countries, we measured the disclosure risk using the Programme for the International Assessment of Adult Competencies (PIAAC) international PUFs for Round 1. PIAAC is an international study that estimates proficiency in literacy, numeracy, and problem solving. The PIAAC survey was conducted for adults 16-65 in the non-institutionalized population, resulting in about 5,000 in-person assessments in each of the 24 countries in Round 1. The PIAAC data dissemination process included a review by the countries for the purpose of data coarsening (e.g., recoding) and variable suppression. It was assumed that each country would follow their nationally-established guidelines and the process was done fairly independently among countries. Some countries have very strict rules that caused some variables to be suppressed, while others with no strict rules allowed all data to be released according to Organisation for Economic Co-operative Development (OECD) plans for the international PUF. After the country review was completed, the international public use file (PUF) for each country was created. The ten PUF variables used in the risk measure illustration using PIAAC data were (indication of the number of levels in parenthesis): Region (Territorial Level 2 (TL2)[1] classification), Age (single year, and 5-year intervals), Education attainment (6 categories), Number of people in the household (top-coded at 6), Living with spouse or partner (2), Children present (2), Sex (2), Born in country (2), Computer experience (2), Native speaker (2). The variables Region, Age (single years), and Native speaker were not released by every country. Native speaker was available for all but one of the countries. It should be mentioned that the U.S. TL2 classification is state, which was too small geography for the sample design. Therefore, Census Region (4) was available in the U.S. national PUF, and was used as the Region variable in the risk measure computations. Lastly, one country was excluded from the analysis since their data was not included in the international PUF due to the national confidentiality rules. The following five scenarios, or combinations, of key variables were evaluated.

- Without Region and Age group
- Including Age group but not Region
- Including both Age group and Region
- Including Age (single years) but not Region
- Including both Age (single years) and Region

In this illustration, from the ETA approach we show the percentage of records that are sample unique. The percentage of records that are sample unique was computed from up to 5-way tabulations among the 10 variables. The ETA risk score based on the number of violations of the threshold rule was also generated. For the SUDA approach, the average score was computed, which is the sum of SUDA scores across all cases divided by the total number of cases. For the Mu-Argus measure, the average score was computed, which is the sum of the probabilities of re-identification divided by the total number of cases. The log-linear measure was not included in this illustration due to excessive computation processing time.

Figure 1 shows the results from the ETA approach in terms of the percentage of records that are sample unique on the y-axis for the countries on the x-axis. The chart shows the impact of including Age and Region in the PUF. Starting from the bottom, the orange marks are the risk levels when Age groups and Region are not included on the PUF. The percent sample unique ranges from 0 to 6%. The blue marks show the risk results when Age groups are included. The percent sample unique range jumps up to 5 to 20%. Next we discuss other combinations of Age and Region. The broken lines are due to the variables not being on the PUF for some countries. The green marks are when Age (single years) but not Region is included. The percent sample unique increases quite a bit when going from Age groups in blue, to Age (single years), in green. Next, at about the same level the red marks are when both Age groups and Region are included. And the highest risk, according to this measure, ranging from about 50% to 85%, is when Age (single years) and Region are included. By using an unweighted threshold rule, the ETA approach ignored the population size, and only considered the sample when computing the risk. Figure 2 shows the average SUDA score by country. The results show similar patterns to the ETA percentage sample unique results, albeit different in relative magnitude between the inclusion/exclusion of sets of variables. We mention that SUDA

[1] See OECD (2013).

was not processed successfully for a small number of countries since they each have one or more very small domains, which causes an error in SUDA. The error occurred when all the cases were identified as unique even before reaching the specified max dimension. There are two possible solutions: 1) drop the small domains, or 2) combine the small domains with others. Although not shown, the ETA risk score based on the number of violations of the threshold rule has similar patterns and relative magnitudes to the average SUDA score between sets of variables released. Li and Krenzke (2013) also confirm with very similar results. A limitation of the percentage of sample unique measure is that the risk rises as the sample size is smaller, even though the sampling rate may be a smaller. Re-identification risk measures attempt to address this issue.

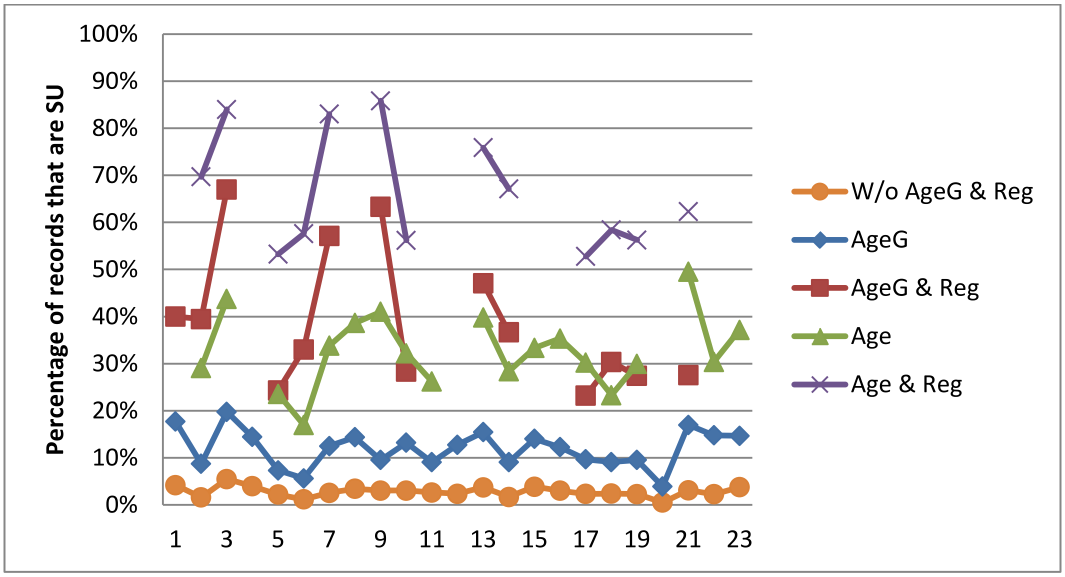

Figure 1. Percentage of records that are sample unique from ETA, by country

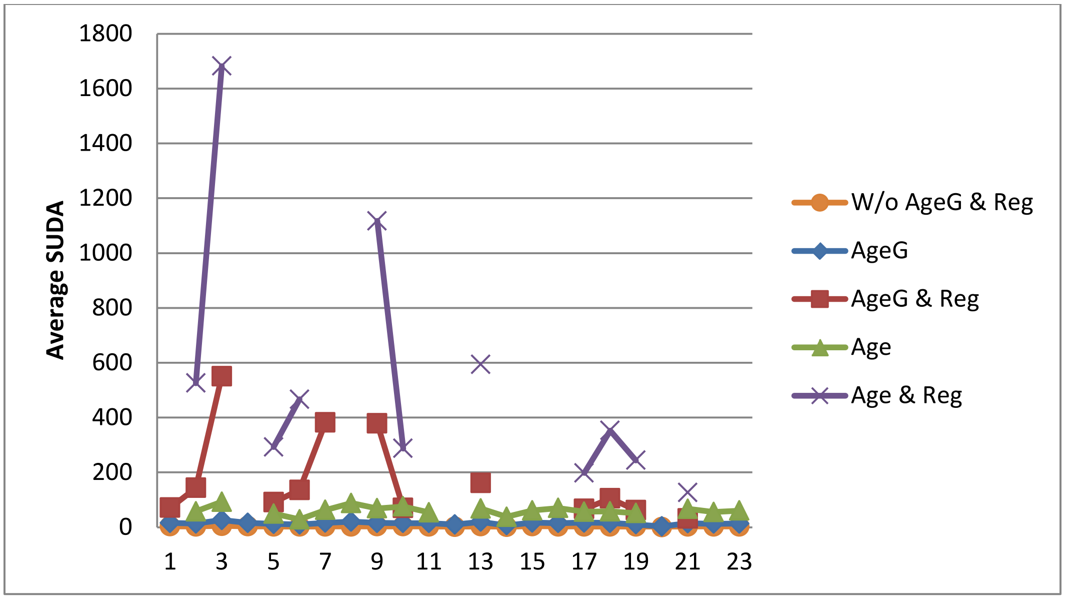

Figure 2. Average SUDA score, by country

The Mu-Argus measure identifies sample uniques among the full cross-classification of the 10 key variables, and then considers whether or not the sample unique is re-identifiable in the population. The expected percentage re-identifiable from the Mu-Argus approach is shown in Figure 3. The results from the Mu-Argus re-identification approach do not follow the same pattern as the ETA and SUDA approaches. In both Figures 2 and 3, the countries are in the same order, sorted by the size of the Mu-Argus measure. The results and conclusion differ depending on the approach. In general, the countries on the right hand side have smaller population sizes than countries on the left hand side. One could use the results to establish a risk threshold to provide guidance to countries as to what to include in the PUF. For example, it might be reasonable to have a threshold of 1%. For the three data points that exceed the threshold, a recommendation based on the threshold rule would be to suppress Age (single years) for two of the countries, and suppress both Age (single years) and Region for the other country.

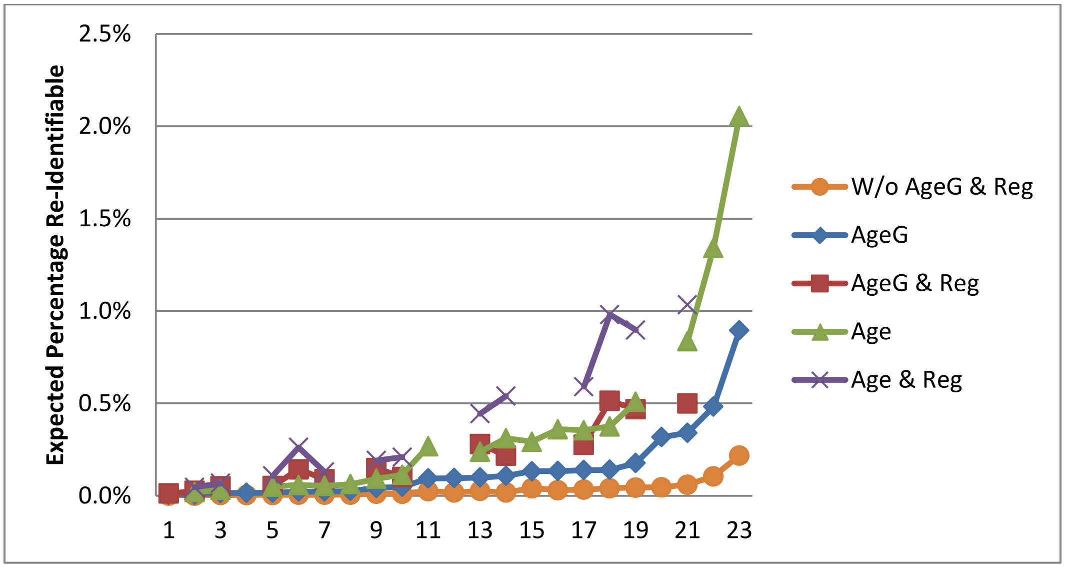

Figure 3. Expected percentage of re-identifable units from Mu-Argus, by country

4. Impact of Missing Values on Risk Measure Estimates

The PIAAC data helped to illustrate the risk measures and the impact of including or excluding certain variables. While the risk computations were being processed, an understanding of how the missing values were treated was needed. For the ETA approach, for each table generated, a complete case analysis is used to determine the sample unique status for each record. For the SUDA, the missing values are treated as a nonmissing category. For the re-identification risk approaches, for Mu-Argus, the missing values are excluded in our application of formulas. For the log-linear modeling approach, the missing values are excluded. Missing values can be treated as a nonmissing category in all approaches if the missing values are set to nonmissing values, like a 7, 8 or 9.

Figure 4 helps to present a simple illustration of the impact of missing values on risk measures. Suppose there are two observations. As shown in the top part of Figure 4, the first record has values for items A, B, and C as 1, 1, and 1, respectively. The 2nd observation has values of A=1, B=1, and a missing value for C. All risk measures discussed assume the missing value is not a value equal to 1, and they either treat the missing value as a separate category, or exclude the data record from the risk computation. The estimated risk for observation 1 depends on the missing value in observation 2. For example, if C = 0 or 2, then observation 1 is unique. If C = 1, then observation 1 is not unique. Therefore, we hypothesize that the existence of missing data inappropriately elevates the estimated risk measure values for the nonmissing cases like observation 1.

Obs	**A**	**B**	**C**
1	1	1	1
2	1	1	.

Obs	**A**	**B**	**C**
1	1	1	1
2	1	1	**0,1,2**

Figure 4. Illustration of the impact of missing values on risk measures

A simulation was conducted in order to look for an approach to improve the estimate of the risk. The base file for the simulation is a fully complete data matrix, consisting of 4,892 records from the U.S. PIAAC PUF with nonmissing data among the 10 key items. The simulation population represents the full sample with all data non-missing. Next, missing values were created for about 20% of the population for the number of persons in the household. Different missing data mechanisms were applied, including missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). In this manner, three sets of 1,000 simulated files of 4,892 records were generated, each differing by the missing data mechanism, and the cases with missing values for the number of persons in the household. For the MCAR mechanism, a simple random sample was selected to generate the missing values. For the MAR mechanism, probabilities were assigned to each records as follows: $p = 1/(1 + e^{(0.35x_1)}$, where x_1 = education attainment. Then for each record, if a uniform random number between the values of 0 an 1 was less than p, then a missing value was created. For the NMAR mechanism, $p = 1/(1 + e^{(0.55y)}$, where y = number of persons in the household.

After the missing values were created, the number of persons in households was temporarily imputed for the sole purpose of measuring the risk. Five imputations were generated using SAS Proc MI for each of the 1000 samples to provide intruder guesses at the true values, while also capturing the uncertainty due to imputation. An ordinal logistic regression imputation model was chosen, and the predictor variables were among the variables used in the PIAAC risk measure computation, namely, Age groups, Education attainment, Sex, Region, and Born in Country.

The schematic in Figure 5 helps to explain the simulation. The risk was computed among all cases and then summarized for the completed (nonmissing) cases (shown in green), which is what is provided in a PUF. On the left is the truth, that is, the full sample with all non-missing data. The left hand side has two pieces shown, 1) the cases that will remain complete, and 2) the cases that will be made missing in the simulation. The middle part of Figure 5 displays the sample with simulated missing values, representing the released PUF. The part on the right in Figure 5 displays the sample with simulated imputed values, which can be thought of as the intruder's guess. We simulated the intruder's guess five times in order to gauge the uncertainty in the guesses.

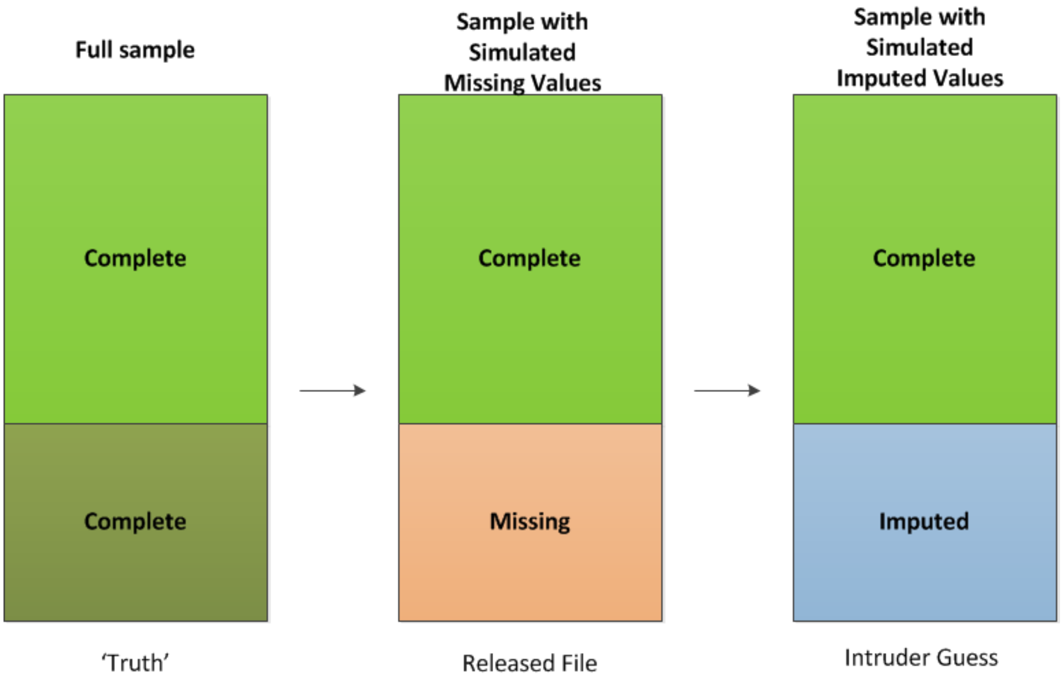

Figure 5. Schematic for simulation design

The comparison among the completes, that is, the nonmissing records (or the green parts in Figure 5). Table 1 shows results for the ETA's percentage of the sample that are sample unique for each of the three missing data mechanisms. For MCAR, the percentage sample unique for the full sample is 21.6%. After missing values were created, the percentage sample unique averaged across the 1000 samples was 25.7%, which is significantly higher (18.7% in relative terms) than the true risk. Among the imputed files the percentage sample unique (21.5%) was only 0.5% (in relative terms) lower than the true risk. Very similar results are shown for the other missing types of MAR and NMAR.

Table 1. Percentage sample unique, by missing data mechanism

Missing		*Percentage Sample Unique*	
Type	*Full sample*	*Missing*	*Imputed*
MCAR	21.6%	25.7%*	21.5%
MAR	21.8%	25.6%*	21.7%
NMAR	21.5%	25.8%*	21.3%

*significantly different

The results shown in Table 2 for the Mu-Argus measure on the number of expected re-identified cases is very similar to the ETA results. For MCAR, the expected number re-identifiable for the full sample is 1.9. After missing values are created, the average expected number re-identified was significantly higher (2.3) than the true risk. Among the imputed files the expected re-identified is 1.9, the same as the true risk. As seen in the ETA results, very similar results are shown for the other missing types of MAR and NMAR.

Table 2. Expected number re-identified, by missing data mechanism

Missing	*Expected Number Re-Identified*		
Type	*Full sample*	*Missing*	*Imputed*
MCAR	1.9	2.3 *	1.9
MAR	1.9	2.2 *	1.9
NMAR	1.8	2.2 *	1.8

*significantly different

This simulation provides indications that the risk measures are likely to overestimate the true risk when missing values are present. The simulation shows a successful correction by temporarily imputing for the variables used in the risk assessment prior to computing the disclosure risk measure. It is also interesting to see that the missing value mechanism had only a slight impact under this simulation set up.

5. Conclusions

In conclusion, we evaluated the disclosure risk impact of certain variables being included or excluded for 23 countries using PIAAC PUF using disclosure risk measures. Different risk measure approaches yield different results and conclusions. As a guide, if it is unknown who is in the sample, then one of the re-identification risk measures (Mu-Argus or the loglinear model approach) is recommended, otherwise, either the ETA or SUDA approach can be useful to identify sparse combinations of variables within the sample. It was determined that a risk threshold could be used to give guidance about what variables should or should not be included, as illustrated with the Mu-Argus measure. In a similar manner, agencies could use risk measures to re-assess their current confidentiality rules or set risk thresholds for their studies.

The impact of missing data on disclosure risk provided indications through simulation that the risk measures overstate the true disclosure risk (under the simulation assumptions). The simulation showed that imputing the variables prior to computing the disclosure risk corrected the overstatement. This emulates a scenario where missing values were not imputed already. If already imputed for the purpose of dissemination then the imputed values can be used in the risk computation. Consideration for dropping the imputation flags from the PUF would reduce the risk further, since imputation without knowing which values were imputed may then be considered as having similar risk-reducing effects as random perturbation.

References

Benedetti, R. and Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. Pre-proceedings of *New Techniques and Technologies for Statistics*, *Vol. 1*. Sorrento, pp. 225–232.

Domingo-Ferrer, J., Torra, V., (2001) A quantitative comparison of disclosure control methods for microdata, Confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies. Doyle, P.; Lane, J.I.; Theeuwes, J.J.M.; Zayatz, L.V. eds., Elsevier, pp. 111-133.

Duncan, G.T. and Lambert, D. (1989). The risk of disclosure of microdata. *Journal of Business and Economic Statistics* 7, 207-217.

El Emam, K., Dankar, F. K., Vaillancourt, R., Roffey, T., and Lysyk, M.(2009). Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian journal of hospital pharmacy*, 62(4):307.

Elliot, M. J., Manning, A. M., and Ford, R. W. (2002). A computational algorithm for handling the special uniques problem. Intern*ational Journal of Uncertainty, Fuzziness and Knowledge Based System*, Vol 10, No. 5, pp 493–509.

Elliot, M.J., Skinner, C.J., and Dale, A. (1998). Special unqiues, random uniques and sticky popilations: Some counterintuitive effects of geographical detail on disclosure risk. *Research in Official Statistics*, 1(2).

Jaro, M.A. (1989) ‘Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida’, Journal of the American Statistical Association, 84, pp.414-20.

Li, J., and Krenzke, T. (2013). Comparing approaches that are used to identify high-risk values in microdata. Census Statistical Disclosure Control Research Project 3. Final report. Washington, DC: U.S. Census Bureau.

OECD (2013). OECD Regions at a glance. Organisation of Economic Cooperation Development. Annex 1. Found in http://www.oecd-ilibrary.org, accessed September 1, 2014.

Polettini, S. (2003). Some remarks on the individual risk methodology. Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Luxembourg.

Privacy Act (1974). The Privacy Act of 1974. U.S. Department of Justice. http://www.justice.gov/opcl/privacy-act-1974 (accessed August 23, 2014).

Skinner, C.J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-linear Models. *Journal of American Statistical Association*, 103, 989–1001.

Winkler, W. (1993) Matching and Record Linkage. U.S. Census Bureau. https://www.census.gov/srd/papers/pdf/rr93-8.pdf (accessed July 17, 2013).