The Survey Statistician, 2022, Vol. 86, 13-21

Ask the Experts

How to Measure Disclosure Risk in Microdata?

Natalie Shlomo[1]

[1] Social Statistics Department, School of Social Sciences, University of Manchester, United Kingdom, natalie.shlomo@manchester.ac.uk

Abstract

In this article we answer the question on how to measure disclosure risk in microdata. We distinguish between two types of microdata: (1) microdata released from social surveys that have undergone statistical disclosure control methods; (2) synthetic microdata generated from statistical modelling. We define the types of disclosure risks and disclosure risk measures for each type of microdata.

Keywords: survey microdata; risk of re-identification; synthetic data; inferential disclosure; privacy models; disclosure risk measures

1 Introduction

Statistical data that are traditionally released by government agencies include microdata from social surveys and tabular data. For each of these traditional outputs, there have been decades of research on how to quantify disclosure risk, statistical disclosure control (SDC) methods and their impact on data utility. However, with increasing demands for new forms of data at higher resolution, in particular linked hierarchical data and 'open' data initiatives, there are even more pressures on government agencies to broaden access and to provide better solutions for the release of statistical data. Examples of solutions are to generate synthetic data based on models built from the original data or to provide access to data through flexible table builders and remote analysis servers. This has led to intensive research and collaboration between the computer science and statistical communities to develop more formal privacy guarantees under SDC and to adapt more perturbative techniques into the SDC tool-kit.

Synthetic data generation has been proposed as an alternative to standard SDC methods for the release of microdata. Traditional SDC methods aim to suppress and perturb existing datasets and often lead to a large loss in utility and analytical power. Synthetic data takes a different approach as it creates a new dataset having the same statistical properties as the original data but containing no data that directly corresponds to real population units. The idea of synthetic data was first introduced by Rubin (1993), who proposed treating each observed data point as if it were missing and imputing it conditional on the other observed data points using a posterior predictive distribution. The data elements are replaced with synthetic values generated from an appropriate probability model.

Copyright © 2022 Natalie Shlomo. Published by International Association of Survey Statisticians. This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Several samples are drawn from the population to take into account the uncertainty of the model and to obtain proper variance estimates. See also Raghunathan, Reiter and Rubin (2003), Reiter (2005a and 2005b), Drechsler (2011) and references therein for more details on generating synthetic data. The synthetic data can also be implemented on parts of data so that a mixture of real and synthetic data is released (Little and Liu 2003).

Here we focus on calculating disclosure risk measures after the application of statistical disclosure control methods or the generation of synthetic data. This is in contrast to disclosure risk assessment in the Computer Science Literature where privacy guarantees are embedded in the perturbation method via a privacy model. These privacy models assume 'attack' scenarios which informs the parameterization of the privacy models according to thresholds. Examples of privacy models in the computer science literature are: k-anonymity, l-diversity and t-closeness where the parameters are k, l and t:

k-anonymity: The key identifying variables are coarsened within equivalence classes such that there are at least $k-1$ individuals in the equivalence class (Sweeney 2002). Equivalence classes are typically defined from quasi-identifying variables such as sex, age group, place of residence.

l-diversity: Determines how well-represented the values of a sensitive variable are within equivalence classes and that there are at least l well-represented values of the variable.

Entropy l-diversity (Machanavajjhala et al. 2006) is defined as follows: Let $p(EC, c)$ be the probability that a record has a value c for a categorical variable C in equivalence class EC. The entropy is: $H(EC) = -\sum_{c \epsilon C} p(EC, c) \log [p(EC, c)]$

A dataset possesses entropy l-diversity if for each EC the entropy $H(EC) \geq \log(l)$.

t-closeness (Li, et al. 2007): Requires the distribution of values of a sensitive variable within equivalence classes to be close (up to t) compared to the univariate distribution of the sensitive variable in the whole dataset.

More on these privacy models can also be found in Domingo-Ferrer, et al. (2008) and Xiao et al. (2010).

Another privacy model gaining much traction in the statistical community is differential privacy (Dwork et al. 2006). A 'worst case' scenario is allowed for, in which the potential intruder has complete information about all the units in the database except for one unit of interest. The definition of a perturbation mechanism M satisfies ε-differential privacy if for all queries on neighbouring databases a and a' differing by one individual and for all possible outcomes defined as subsets $S \in Range(M)$ we have: $p(M(a) \in S) \leq e^{\varepsilon} p(M(a') \in S)$.

This means that observing a perturbed output S, little can be learnt (up to a degree of e^{ε}) and the intruder is unable to decipher whether the output was generated from database a or a'. In other words, the ratio $p(M(a) \in S) / p(M(a') \in S)$ is very small (at most e^{ε}). The solution to guarantee differential privacy in the computer science literature is by adding noise/perturbation to the outputs of the queries under specific parameterizations based on the privacy budget ε and the sensitivity of the query, which is the maximum difference in the possible output of a query with and without the presence of a single individual.

In Section 2, I discuss the types of disclosure risks for microdata. In Section 3, I describe how to estimate a disclosure risk measure to assess the risk of re-identification in disclosure-controlled survey microdata. In section 4, I describe disclosure risk measures that can be used after the generation of synthetic data to assess attribute and inferential disclosure. I close with a conclusion in section 5.

2 Types of Disclosure Risks for Microdata

In the SDC literature, we define the notion of an 'intruder' as someone who wants to attack statistical data for malicious intent and cause a breach in confidentiality. Two main disclosure risks are: (1)

identity disclosure where a statistical unit can be identified based on a set of cross-classified quasi-identifying variables that are typically categorical, such as age, gender, occupation and place of residence; (2) attribute disclosure where new information can be learnt about an individual or a group of individuals. Disclosure risk scenarios form the basis of possible means of disclosure, for example, the ability of an intruder to match a dataset to an external public file based on a common set of quasi-identifying variables; the ability of an intruder to identify unique individuals through visible and rare attributes; the ability of an intruder to difference nested tables and obtain small counts; and the ability of an intruder to form coalitions with other intruders.

For the release of survey microdata that are disseminated from social surveys, the main concern is the risk of re-identification since this is a prerequisite for individual attribute disclosure where many sensitive variables such as income or health outcomes, can be revealed following an identification. Naturally, sampling from the population provides a priori protection since an intruder cannot be certain whether a sample unique, i.e. a unit that is unique in the sample with respect to some quasi-identifying variables, is a population unique. Note there is an implicit assumption of no 'response knowledge' meaning that the intruder does not know who was drawn into the sample of the survey.

Inferential disclosure is another type of disclosure risk that is becoming more prominent with the ongoing research and development into web-based interactive data dissemination. Inferential disclosure risk is the ability to learn new attributes with high probability and thus is a more general form of individual and group attribute disclosure and the terms are often used interchangeably. For example, datasets can be manipulated and combined in such a way that there is a high prediction power between variables in the dataset or combinations of data releases that can be differenced to reveal individual data points. Attribute disclosure and the more general inferential disclosure are particularly relevant for assessing disclosure risks in fully synthetic data. This is because there is a break in the link between quasi-identifying and sensitive variables in a fully synthetic dataset, but it may still be possible to disclose sensitive information about groups of individuals.

3 Quantifying the Risk of Re-identification in Survey Microdata

The basic definition of the risk of re-identification is the probability of correctly being able to match the survey microdata with a unit in the population. If the characteristics of the population are known, such as measured in a population register or census, this probability would be relatively straightforward to calculate. However, this is rarely the case since within government agencies, samples are often drawn from area or address-based sample frames. A statistical modelling framework is then needed to estimate the probability of re-identification. This probability is conditional on the released data and information available to the intruder and defined with respect to a probabilistic model and assumptions about how the data is generated (knowledge of the sampling process). The model is based on the set of quasi-identifiers available to the intruder and available in released data which, when cross-classified for the released data, form a contingency table that can be used to identify cells with small sample sizes, and we particularly focus on the sample uniques. The risk of re-identification is based on the notion of population uniqueness in the contingency table: given an observed sample unique, what is the probability that the cell is also a population unique?

The probabilistic modelling to estimate population uniqueness from the observed survey microdata was developed under two approaches: a fully model-based framework taking into account all of the information available to intruders and modelling their behaviour (Duncan and Lambert 1989, Lambert 1993 and later Reiter 2005c) and a more simplified approach that restricts the information that would be known to intruders (Bethlehem, et al. 1990, Benedetti, et al. 1998, Skinner and Holmes 1998, Elamir and Skinner 2006).

Individual per-record risk measures in the form of a probability of re-identification are estimated. These per-record risk measures are then aggregated to obtain global risk measures for the entire file. Denote by F_k the population size in cell k of a table spanned by quasi-identifying variables having K cells and by f_k the sample size. We have $\sum_k F_k = N$ and $\sum_k f_k = n$ with N the total population size and n the size of the released sample. The set of sample uniques is defined as: $SU = \{k: f_k = 1\}$

since these are the potential high-risk records with the potential to be population uniques. Two global disclosure risk measures (where I is the indicator function) are the following:

Number of sample uniques that are population uniques: $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$

Expected number of correct matches for sample uniques assuming a random assignment within cell k (the match probability) $\tau_2 = \sum_k I(f_k = 1)\ \frac{1}{F_k}$

We assume that the population frequencies F_k are unknown and need to be estimated from a probabilistic model where the risk measures are then:

$$\hat{\tau}_1 = \sum_k I(f_k = 1)\hat{P}(\ F_k = 1|f_k = 1) \text{ and } \hat{\tau}_2 = \sum_k I(f_k = 1)\hat{E}(\frac{1}{F_k}\ |f_k = 1) \quad (1)$$

Skinner and Holmes (1998) and Elamir and Skinner (2006) propose a Poisson distribution and a log-linear model to estimate disclosure risk measures in (1). In this model, they assume that $F_k \sim Pois(\lambda_k)$ for each cell k. A sample is drawn by Poisson or Bernoulli sampling with a sampling fraction π_k in cell k: $f_k|F_k \sim Bin(F_k, \pi_k)$. It follows that:

$$f_k \sim Pois(\pi_k\lambda_k) \text{ and } F_k|f_k \sim Pois(\lambda_k(1 - \pi_k)) \quad (2)$$

where the population cell counts F_k are assumed independent given the sample cell counts f_k.

The parameters λ_k are estimated using log-linear modeling. The sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k\lambda_k$. A log-linear model for the μ_k is expressed as: $\log(\mu_k) = x_k{}'\boldsymbol{\beta}$ where x_k is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is obtained by solving the score equations:

$$\sum_k(\ f_k - \pi_k \exp(x_k'\boldsymbol{\beta}))x_k = 0 \quad (3)$$

The fitted values are then calculated by: $\hat{\mu}_k = \exp(x_k'\widehat{\boldsymbol{\beta}})$ and $\hat{\lambda}_k = \hat{\mu}_k/\pi_k$. Individual disclosure risk measures for cell k are:

$$P(\ F_k = 1|f_k = 1) = \exp\ (\lambda_k(1 - \pi_k)) \text{ and } E(\frac{1}{F_k}\ |f_k = 1) = (1 - \exp\ (\lambda_k(1 - \pi_k)))/\ (\lambda_k(1 - \pi_k)) \quad (4)$$

Plugging $\hat{\lambda}_k$ for λ_k in (4) leads to the estimates $\hat{P}(\ F_k = 1|f_k = 1)$ and $\hat{E}(\frac{1}{F_k}\ |f_k = 1)$ and then to $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1). Rinott and Shlomo (2007b) consider confidence intervals for these global risk measures.

Skinner and Shlomo (2008) develop goodness-of-fit criteria for selecting the main effects and interactions of the quasi-identifying variables for the log-linear model based on estimating and (approximately) minimizing the bias of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. In addition, they address the estimation of disclosure risk measures under complex survey designs with stratification, clustering and survey weights. While the method described assumes that all individuals within cell k are selected independently using Bernoulli sampling, i.e. $P(f_k = 1|F_k) = F_k\pi_k(1 - \pi_k)^{F_k - 1}$, this may not be the case when sampling clusters (e.g. households). In practice, key variables typically include variables such as age, sex and occupation that tend to cut across clusters. Therefore, the above assumption holds in practice in most household surveys and does not cause bias in the estimation of the risk measures. Inclusion probabilities may vary across strata, the most common stratification is on geography. Strata indicators should always be included in the quasi-identifying variables to take into account differential inclusion probabilities in the log-linear model. Under complex sampling, the λ_k can be estimated consistently using pseudo-maximum likelihood estimation (Rao and Thomas 2003), where the estimating equation in (3) is modified as:

$$\sum_k(\ \hat{F}_k - \exp(x_k'\boldsymbol{\beta}))x_k = 0 \quad (5)$$

and $\hat{F}_k$ is obtained by summing the survey weights in cell k: $\hat{F}_k = \sum_{i \in k} w_i$. The resulting estimates λ_k are plugged into expressions in (4) and π_k is replaced by the estimate $\hat{\pi}_k = f_k / \hat{F}_k$. The goodness-of-fit criteria are also adapted to the pseudo-maximum likelihood approach.

The probabilistic modelling presented here and in other related work in the literature assumes that there is no measurement error in the way the data is recorded. Besides typical errors in data capture, key variables can also purposely be perturbed as a means of masking the data, for example through record swapping or the post randomization method (PRAM) (Gouweleeuw, et al. 1998). Shlomo and Skinner (2010) adapt the estimation of the risk of re-identification to take into account measurement (perturbation) errors. We denote the cross-classified quasi-identifying variables in the population and the microdata as X and assume that X in the microdata have undergone some perturbation error denoted by the value $\tilde{X}$ and determined independently by a misclassification matrix M:

$$M_{kj} = P(\tilde{X} = k | X = j) \tag{6}$$

Under small sampling fractions and small rates of perturbation as reflected in the misclassification matrix in (6), we can assume that only the diagonal of the misclassification matrix is needed, i.e. the probabilities of not being perturbed. The estimate of $\hat{\tau}_2$ in (1) can be obtained by the probabilistic modelling framework described above on the misclassified sample:

$$\hat{\tau}_2 = \sum_k I\left(\tilde{f}_k = 1\right) M_{kk} \, \hat{E}\left(\frac{1}{\tilde{F}_k} \middle| \tilde{f}_k\right) \tag{7}$$

where $\tilde{f}_k$ are the observed perturbed counts and $\tilde{F}_k$ represent population counts.

There have been many other contributions extending the Poisson-log linear modelling framework for estimating the risk of re-identification in survey microdata. Ichim (2008) considers extensions by introducing the survey weights in the analysis of the contingency tables and also proposes a maximum penalized-likelihood approach to obtain smoother estimates of the risk of re-identification. Forster and Webb (2007) extend the log-linear modelling framework to a model averaging approach rather than requiring to choose a model a priori. They use a Bayesian model averaging technique according to several possible log-linear models but limit the models to decomposable geographical models. The posterior distribution under model uncertainty is hence obtained as a weighted average of the posterior distribution under the various models. Rinott and Shlomo (2006 and 2007a) generalize the probabilistic modelling using the Negative Binomial distribution rather than the Poisson distribution and implement the probabilistic modelling framework on local 'neighbourhoods' of the sample uniques. Manrique-Vallier and Reiter (2012) propose an alternative to log-linear models for datasets with sparse contingency tables according to the quasi-identifying variables using a Bayesian version of grade of membership models and they use a Markov Chain Monte Carlo algorithm for fitting the model. Carota, et al. (2015) applied a Bayesian semi-parametric version of log-linear models, specifically a mixed effects log-linear model with a Dirichlet process prior.

A new direction is currently under development to measure the risk of re–identification for non-probability data sources. More specifically, there are registers in the public domain where the membership of the register is not known and is sensitive. Examples of registers are of individuals with a medical condition, such as Cancer or HIV, or registers that include membership to a loyalty card scheme. Shlomo and Skinner (forthcoming) focus on this new setting by extending the framework of probabilistic modelling. The microdata from a random sample can still be used to estimate population parameters under the probabilistic modelling framework for estimating the risk of re-identification, however the complication is that another set of parameters needs to be estimated: the propensities of membership for the individuals in the register. This accounts for the selection bias in the register and the deviation from the general population.

For partially synthetic data, assessing disclosure risk where some values of variables are not changed has been further shown in Reiter and Mitra (2009) and Drechsler and Reiter (2011). There, the authors assume that an intruder knows the values of a single target record and then searches the released data to identify the record. Other work on identity disclosure for fully synthetic data has

been shown in Reiter et al. (2014). The authors assume that an intruder has prior knowledge of the entire dataset except for one record and then attempts to quantify the risk of re-identification using Bayesian estimation to obtain the posterior distributions of confidential data given the released data. The intruder then evaluates the posterior distribution of possible original values for the one unknown record, given the released synthetic data and information about the data generation mechanism and uses values with high probability as reasonable guesses for the unknown true values.

4 Quantifying the Risk of Attribute Disclosure for Synthetic Data

Fully synthetic data should lead to a break between the identifying variables and the sensitive target variables, and hence the main focus for quantifying disclosure risk in fully synthetic data is to measure attribute disclosure (and more generally, inferential disclosure). This disclosure risk is based on being able to infer characteristics of individuals in the datasets, particularly groups of individuals.

With respect to developing disclosure risk measures after the generation of the data, one measure that can be used to identify skewness in the distribution of categories c of a variable C in equivalence class EC is the entropy. The entropy of the distribution obtains a maximum value if the distribution of the categories is uniform and a minimum value if the distribution is degenerate (there is only one category represented). We can transform the entropy defined in Section 1 to the E measure so that we obtain a value between 0 and 1 as follows: $E = 1 - H(EC)/\log(K)$ where K is the number of categories of the variable (Antal et al. 2014). We also define the L measure which measures the percentage of the number of categories of the sensitive variable similar to the principle of l-diversity.

We can develop distance metrics that compare the overall distributions in the original data versus synthetic data for a particular variable and more specifically within equivalence classes EC. Distance metrics include Kullback-Leibler distance, the Total Variation (TV) and Hellinger's Distance (HD). For a categorical variable C in equivalence class EC, the Hellinger's Distance is equal to:

$HD_{EC}(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{c\epsilon C}(\sqrt{p(EC,c)} - \sqrt{q(EC,c)})^2}$ where $p(EC,c)$ is the distribution of C in the original data and $q(EC,c)$ is the distribution of C in the synthetic data. The Total Variation is equal to: $TV_{EC}(P,Q) = \frac{1}{2}\sum_{c\epsilon C}|p(EC,c) - q(EC,c)|$.

Note that these distance metrics can also be used for utility measures, i.e. measures that express the usefulness of the data for statistical analysis, and hence we blur the lines about what constitutes measures of disclosure risk and what measures utility.

Similar to the privacy model of t-closeness, we can use distance metrics comparing the distribution in the synthetic data for variable C in an equivalence class EC with the overall univariate distribution in the original data, denoted $Q(c)$. In this case, the Total Variation is $TV(P,Q) = \frac{1}{2}\sum_{c\epsilon C}|p(EC,c) - Q(c)|$.

Elliot (2014) and Taub et al. (2018) defined the Differential Correct Attribution Probability ($DCAP$) framework. It assumes that the intruder has access to the synthetic data s and has knowledge of an equivalence class denoted $EC_{o,i}$ for individual i in the original dataset o and wants to learn the value of a sensitive variable $T_{s,i}$. The intruder then identifies all the records that match on EC_o in the synthetic data s. If the proportion of records in the equivalence class on $\{EC_s, T_s\}$ is high then the intruder can infer the value $T_{s,i}$ for $T_{o,i}$. In summary, $DCAP$ measures the proportion of records for equivalence class EC_o that have the same target value in the synthetic data as the original value. More formally, define D_o the original data composed of equivalence classes EC_o and sensitive variables T_o: $D_o = \{EC_o, T_o\}$ and similarly, the synthetic data is defined as: $D_s = \{EC_s, T_s\}$. For each individual i we define: $DCAP_{o,i} = \sum_{i=1}^{N} I(T_{oi} = T_{si} \text{ and } EC_{oi} = EC_{si})/\sum_{i=1}^{N} I(EC_{oi} = EC_{si})$ where N is the size of the dataset (assumes the same N in the synthetic and original data) and I is the indicator function taking a value of 1 if the condition is satisfied, otherwise 0. Similarly calculate $DCAP_{s,i}$ in the

synthetic data. Note that it is possible that the denominator in $DCAP_{s,i}$ can be 0 and may be undefined. In that case, we can define the measure as 0. The baseline is: $DCAP_{b,i} = \frac{1}{N}\sum_{i=1}^{N} I(T_{oi} = T_{si})$. The original and baseline measures serve as bounds for comparing the $DCAP_{s,i}$ and ensuring that it is sufficiently reduced.

Chen et al. (2019) noted that this original measure of $DCAP$ is similar to the distance-based utility measures and proposed to adapt the $DCAP$ framework to only those records that are unique in the synthetic data in the EC. The risk measure is defined as Targeted Correct Attribution Probability ($TCAP$).

We can see that there is a clear connection between $DCAP$ and the l-diversity privacy model as the less diverse the sensitive variables in the synthetic data, the higher risk of discovering a sensitive attribute.

5 Conclusion

The framework for measuring the risk of re-identification as discussed in Section 3 based on estimating the probability of population uniqueness is well established although many different approaches have been proposed in the SDC literature to estimate these disclosure risk measures. However, as can be seen in Section 4, disclosure risk measures for synthetic data after its generation are still ad-hoc and a more formal framework is needed for measuring the risk of attribute disclosure. In addition, appropriate software needs to be developed which will enable the framework to be embedded in the SDC tool-kit at government agencies.

References

Antal, L., Shlomo, N. and Elliot, M. (2014) Measuring Disclosure Risk with Entropy in Population Based Frequency Tables. In Privacy in Statistical Databases 2014, (Ed. J. Domingo-Ferrer), Springer LNCS 8744, 62-78.

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) Disclosure Control of Microdata. Journal of the American Statistical Association, **85**, 38-45.

Benedetti, R., Capobianchi, A., and Franconi, L. (1998) Individual Risk of Disclosure Using Sampling Design. Contributi Istat.

Carota, C., Filippone, M. Leombruni, R. and Polettini, S. (2015) Bayesian Nonparametric Disclosure Risk Estimation via Mixed Effects Log-linear Models. Annals of Applied Statistics, **9(1)**, 525 – 546.

Chen, Y., Taub, J. and Elliot, M. (2019) Trade-off Between Information Utility and Disclosure Risk in GA Synthetic Data Generator. Conference of European Statisticians, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 29-31 October 2019, The Hague, the Netherlands. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Chen_Taub_Elliot_AD.pdf

Domingo-Ferrer, J. and Torra, V. (2008) A Critique of k-anonymity and Some of its Enhancements. In 2008 Third International Conference on Availability, Reliability and Security. IEEE, 990-993.

Drechsler, J. (2011) Synthetic Datasets for Statistical Disclosure Control. Lecture Notes in Statistics (LNS) 201, NY: Springer.

Drechsler, J. and Reiter, J. (2011) An Empirical Evaluation of Easily Implemented, Non-parametric Methods for Generating Synthetic Data. Computational Statistics and Data Analysis, **55**, 3232-3243.

Duncan, G. and Lambert, D. (1989) The Risk of Disclosure for Microdata. Journal of Business and Economic Statistics, **7**, 207-217.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating Noise to Sensitivity in Private Data Analysis. In Theory of Cryptography TCC (eds. S. Halevi and R. Rabin). Heidelberg: Springer, LNCS 3876, 265-284.

Elamir, E. and Skinner, C.J. (2006) Record-Level Measures of Disclosure Risk for Survey Micro-data. Journal of Official Statistics, **22**, 525-539.

Elliot, M. (2014) Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team. Available at: https://tinyurl.com/syllsDR

Forster, J.J. and Webb, E.L. (2007) Bayesian Disclosure Risk Assessment: Predicting Small Frequencies in Contingency Tables. Journal of Royal Statistical Society Series C, **56**, 551–570.

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P. (1998) Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, **14**, 463-478.

Ichim D. (2008) Extensions of the Re-identification Risk Measures based on Log-linear Models. In Privacy in Statistical Databases (eds. J. Domingo-Ferrer and Y. Saygin), Lecture Notes in Computer Science 5262. Springer, Berlin, 203-212.

Lambert, D. (1993) Measures of Disclosure Risk and Harm. Journal of Official Statistics, **9**, 313-331.

Li, N., Li, T. and Venkatasubramanian, S. (2007) t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering.

Little, R.J.A., and Liu, F. (2003) Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 6. http://www.bepress.com/umichbiostat/paper6

Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M. (2006) l-diversity: Privacy Beyond k-anonymity. In 22nd International Conference on Data Engineering (ICDE'06), IEEE, 24.

Manrique-Vallier, D. and Reiter, J. P. (2012) Estimating Identification Disclosure Risk Using Mixed Membership Models. Journal of the American Statistical Association, **107**, 1385–1394.

Raghunathan, T.E., Reiter, J. and Rubin, D. (2003) Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, **19(1),** 1-16.

Rao, J. N. K., and Thomas, D. R. (2003) Analysis of Categorical Response Data from Complex Surveys: An Appraisal and Update. In Analysis of Survey Data (eds. R. L. Chambers and C. J. Skinner), Wiley, Chichester, UK, 85–108.

Reiter, J.P. (2005a). Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. Journal of the Royal Statistical Society A, **168(1),** 185–205.

Reiter, J. (2005b) Using CART to Generate Partially Synthetic, Public Use Microdata. Journal of Official Statistics, **21,** 441–462.

Reiter, J.P. (2005c) Estimating Risks of Identification Disclosure in Microdata. Journal of the American Statistical Association, **100**, 1103-1112.

Reiter, J. and Mitra, R. (2009) Estimating Risks of Identification Disclosure in Partially Synthetic Data. Journal of Privacy and Confidentiality, **1(1)**, 99–110.

Reiter, J., Wang, Q. and Zhang, B. (2014) Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. Journal of Privacy and Confidentiality, **6(1)**, 17–33.

Rinott, Y. and Shlomo, N. (2006) A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation. In Privacy in Statistical Databases (eds. J. Domingo-Ferrer and L. Franconi), Lecture Notes in Computer Science 4302 Springer, Berlin, 82–93.

Rinott, Y. and Shlomo, N. (2007a) A Smoothing Model for Sample Disclosure Risk Estimation. In Complex Datasets and Inverse Problems (eds. R. Liu, W. Strawderman and C.-H. Zhang), Institute of Mathematical Statistics Lecture Notes, Monograph Series 54, Ohio, 161-171.

Rinott, Y. and Shlomo, N. (2007b) Variances and Confidence Intervals for Sample Disclosure Risk Measures. In Bulletin of the International Statistical Institute: Proceedings of the 56th Session of the International Statistical Institute, ISI'07, Lisbon, 1090–1096.

Shlomo, N. and Skinner, C.J. (2010) Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. Annals of Applied Statistics, **4(3)**, 1291-1310.

Shlomo, N. and Skinner, C.J. (forthcoming) Measuring Risk of Re-identification in Microdata: State-of-the Art and New Directions, Journal of the Royal Statistical Society, Series A.

Skinner, C.J. and Holmes, D. (1998) Estimating the Re-identification Risk per Record in Microdata. Journal of Official Statistics, **14**, 361-372.

Skinner, C. J. and Shlomo, N. (2008) Assessing Identification Risk in Survey Micro-data Using Log Linear Models. Journal of American Statistical Association, **103(483),** 989-1001.

Sweeney, L. (2002) k-anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, **10(5)**, 557-570.

Taub J., Elliot M., Pampaka M. and Smith D. (2018) Differential Correct Attribution Probability for Synthetic Data: An Exploration. In: (eds. Domingo-Ferrer J., Montes F.) Privacy in Statistical Databases. PSD 2018. Lecture Notes in Computer Science, Vol 11126. Springer.

Xiao, X., Yi, K. and Tao, Y. (2010) The Hardness and Approximation Algorithms for l-diversity. In Proceedings of the 13th International Conference on Extending Database Technology, ACM, 135-146.