

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

Synthetic Data Disclosure Risk Assessment: A Literature Review

by Zhe Si Yu

Release date: September 8, 2025



Synthetic Data Disclosure Risk Assessment: A Literature Review

Zhe Si Yu¹

Abstract

The adoption of synthetic data generation as a confidentiality measure is increasing in statistical agencies worldwide, including at Statistics Canada. This approach provides an alternative to the traditional dissemination of anonymized public microdata files, offering both privacy protection and data utility. However, the creation of synthetic data presents challenges in assessing and mitigating disclosure risks. This paper reviews the different types of disclosure risks, that being attribute, membership and identity disclosure, and presents some of the associated methods for measuring risk. The paper presents prominent risk assessment metrics and discusses practical methods for disclosure control in data synthesis. Methods for assessing disclosure risks usually produce a metric that can be used to gauge the risk, but there is little consensus on threshold values for these metrics. It is also important to focus on importance of balancing utility and confidentiality, which needs further discussion in context of these methods. The paper concludes by offering insights and recommendations about managing disclosure risk while creating synthetic data as well as providing some ideas on future directions for research and practical implications for managing disclosure risks in synthetic data.

Key Words: Synthetic data; Disclosure risk assessment; Confidentiality.

1. Introduction

Synthetic data has garnered significant attention in data privacy due to its potential to offer high-utility data while preserving confidentiality. The definition of synthetic data used for this paper is any dataset generated through a synthesis model that was trained on real data that mimic the real dataset statistically. A good synthetic dataset should generate similar statistical conclusions as the real data. National Statistical Organization (NSOs) like Statistics Canada increasingly turn to synthetic data as an alternative to traditional anonymization (Gauvin, 2021; Sallier, 2021) e.g. suppression. Currently, Statistics Canada primarily uses synthetic data for training, educational and simulation purposes, never for official statistics. Statistics Canada has previously contributed to the official UNECE guide for synthetic data generation (UNECE, 2022), which detailed some synthetic data disclosure risk assessment. This paper is an extension of our previous work to investigate the current state of research for synthetic data disclosure risk assessment.

In the context of this paper we define fully and partially synthetic data; fully synthetic datasets replace all real records with synthetically generated data, while partially synthetic datasets generally retain the real records, substituting only the sensitive attributes with synthesized values. However, the scope of this paper will focus on fully synthetic data.

Despite the advantages of synthetic data, they are not devoid of all disclosure risks. It is important to have clear metrics for measuring the risks of synthetic data, especially if we are to see more practical use of it. This literature review explores the three primary disclosure risks—identity, attribute, and membership disclosure—and reviews techniques to assess these risks, balancing the trade-offs between data utility and disclosure risk².

¹Zhe Si Yu, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6, zhesi.yu@statcan.gc.ca

²The content of this paper reflects the positions of the author, not necessarily that of Statistics Canada.

2. Types of Disclosure Risk

2.1. Identity Disclosure

Identity disclosure risk occurs when an intruder associates a known individual with a released data record (Reiter, 2005). It is widely accepted in the literature that fully synthetic data has a low identity disclosure risk, as there is no one-to-one correspondence between the records in the fully synthetic data and those in the real dataset (Hu, 2021; Reiter, 2005; Reiter & Mitra, 2009). Intuitively, because all records in a fully synthetic dataset are generated, there is no meaningful linkage to the real dataset.

2.2. Attribute Disclosure

Attribute disclosure risk in synthetic data is measured by estimating probabilities for exposing the real target value of records in the original data (Hittmeir et al., 2020; Hornby & Hu, 2021). This scenario assumes that the intruder has some prior knowledge of the original (real) dataset and/or knowledge of the synthesizer. Attribute disclosure spans a spectrum of concern, as it depends on whether the attribute itself is sensitive (and needs protection) or not. This decision often depends on the dataset.

In general, most techniques for assessing attribute disclosure risk involve calculating a score based on how likely it is, given previous knowledge, that an attacker can accurately guess a record's attribute value. Generative models used for synthetic data replicate correlations between attributes by constructing a conditional generative process. For example, the Fully Conditional Specification method used by Synthpop in R (Nowok et al., 2016) generates each attribute (except the first) sequentially by conditioning on the attributes previously generated. The conditional nature of these processes has led most researchers to adopt Bayesian estimation methods (Hornby & Hu, 2021; Hu et al., 2014; Hu, 2021; Reiter et al., 2014) for risk quantification.

2.3 Membership Disclosure

Membership disclosure risk is especially relevant for synthetically generated datasets. Membership disclosure occurs when a record in the population is identified as being a member of the real dataset through the release of the synthetic data and/or the training model. We note that there is a difference in approach for membership disclosure between synthetic data and the well-explored, related topic of generative models. Generative models are in the more general category into which synthetic models fall, where it includes any statistical models that take some data as input and generates some desired data as output. Most techniques explored for generative models assume a black box model that is nonetheless available for attackers, whereas the model is usually not available for synthetic data (Zhao et al., 2019). Since the assumptions of prior knowledge are different between generative model and synthetic data disclosure, we cannot directly apply previous research of generative models onto synthetic data. Synthetic data disclosure assessment needs their own methods (Zhao et al., 2019).

2.4 Perceived Risk

The purpose of synthetic data is to accurately represent its original data statistically. However, a good replication will inevitably appear to be suspiciously real, often appearing like an identity disclosure, creating a unique issue called perceived risk. While on a fully synthetic dataset, all records are fictitious, they may coincidentally be generated to be very similar to a real record, enough to cause the perception that real information has been disclosed. This is not a real disclosure risk, but nonetheless an issue that needs to be tackled.

3. Disclosure Risk Assessment Methods

Specific methods in assessing the disclosure risks of attribute and membership disclosure will be discussed below. As stated above, the risk of identity disclosure with fully synthetic data is low and perceived risk is not a real disclosure risk. As such, we will only discuss assessment methods for attribute and membership disclosure.

3.1. Attribute Disclosure

Attacks that aim for attribute disclosure are referred to as the *Attacker's Classification Problem* (Hittmeir et al., 2020; Hornby & Hu, 2021). The attacker, equipped with some background knowledge, the synthetic dataset(s) and the values of the key attributes of some record in the original dataset, seeks to predict the target value of some record.

As described in Reiter *et al.* (Reiter et al., 2014), consider an original dataset consisting of microdata with n records and m attributes, where $D = \{(x_i, y_i) : i = 1, \dots, n\}$ is the dataset, x_i is the vector of the i -th record's values of non-sensitive attributes, and y_i is the vector of the i -th record's values of sensitive attributes which are subject to synthesis. Note that for fully synthetic data, $X = \{x_i, i = 1, \dots, n\}$ is empty where all variables will be synthesized. We also assume that the data providers release m copies of synthetic datasets, $Z = \{Z^{(1)}, \dots, Z^{(m)}\}$. We assume that the attacker wants to learn the attribute y_i for some real record i , where $Y = \{y_i, i = 1, \dots, n\}$. Let $B = \{A, S\}$ be the background knowledge of the attacker, where A is information known about the original data and S is knowledge about the synthesizer (Reiter et al., 2014). Then for a guess y^* for y_i , the attacker recovers the Bayesian posterior distribution,

$$P(Y_i = y^* | Z, X, A, S) = \frac{P(Z | Y_i = y^*, X, A, S)P(Y_i = y^* | X, A, S)}{\sum_y P(Z | Y_i = y, X, A, S)P(Y_i = y | X, A, S)} \quad (1)$$

The data provider is then able to use the resulting value to compute several risk measures for the released synthetic dataset(s). One method is to compute

$$R_i = [\operatorname{argmax}_{y^*} P(Y_i = y^* | Z, X, A, S) = y_i], \quad (2)$$

then decide whether $R = \sum_{i=1}^n R_i/n$ is acceptably low (Reiter et al., 2014). Intuitively, the attacker wants the maximum probability, among all y_i , that their guess, y^* , would equal to y_i , based on prior information of the data, records and model. It is recommended to maximize A i.e. assume that the attacker knows all y_i except the one being attacked. It has been noted that this assumption is overly conservative, since in many cases this level of knowledge would be unrealistic (Reiter et al., 2014). Nonetheless, Reiter *et al.* showed that their method can be simulated and tested with different assumptions, (Reiter et al., 2014) meaning that we can compare the changes in R values.

While Bayesian estimation remains the more popular choice for assessing attribute disclosure (Hornby & Hu, 2021; Reiter et al., 2014), there are other methods. For the concept of *Correct Attribution Probability* (CAP) and its derivatives (Hittmeir et al., 2020), it is also assumed that the attacker knows the values of a set of key attributes for an individual in the original dataset, and wants to learn some unknown, target attribute. This method does not assume that the attacker also has access or knowledge of the synthetic model itself, which sets itself different from the Bayesian estimation techniques mentioned above. The formula for the CAP score is

$$CAP_{\blacksquare, j} := P_{\blacksquare}(T_{o, j} | K_{o, j}) = \frac{\sum_{i=1}^n [T_{\blacksquare, i} = T_{o, j} \wedge K_{\blacksquare, i} = K_{o, j}]}{\sum_{i=1}^n [K_{\blacksquare, i} = K_{o, j}]}, \quad (3)$$

Where $j \in \{1, \dots, n\}$, $K_{\blacksquare, j}$ is a vector representing the values of the key attributes of the j -th record in the original dataset, $T_{\blacksquare, j}$ is the corresponding value of the target attribute, and $\blacksquare = \{o, s\}$ is the original and synthetic data, respectively (Hittmeir et al., 2020). The idea of this attack is that the intruder can search over all records in the synthetic data that match the key attribute values known by them. Then, within these records, the attackers can calculate the distribution of the occurring values of the target attribute. The mean value of CAP_s can then be compared to CAP_o .

Since Bayesian techniques are more conservative in some regard with all its assumptions, while *CAP* does not have any assumptions, we can use both families of technique to set an upper and lower bound of attribute risk. In more extreme situation we can use *CAP* to estimate the attribute disclosure risk for the original data, which can then be set as the highest bound for risk.

3.2. Membership Disclosure

One method of attack is called the partitioning method: when an adversary has an attack dataset that is a sample from the same population as the real dataset that is assumed to be used in training for the synthesis model (Zhang et al., 2022). The attacker matches records from their sample with the synthetic data, and membership disclosure occurs when a matching record is also in the training (real) set. Membership disclosure risk assessment thus aims to quantify the difference between synthetic and real data. We define some relevant statistical terms that are commonly used in risk assessment. Precision is defined as the proportion of true positives in all positives, and recall is the proportion of true positives in true positives and false negatives.

In order to estimate the risk underlying such an attack, the data custodian can separate the real data into training and holdout data. We assume that the attacker has complete information on m records, where mt are drawn for the training sample and $m(1 - t)$ are drawn from the holdout sample. $t \in (0,1)$ is a chosen proportion of records chosen to be in the training sample. The default for t selected in the paper is 0.5 (El Kababji et al., 2023). The synthesis model is then trained on the training data, producing synthetic data that can be compared with the real data (both training and holdout). We can then calculate the minimum distance between every record in the real data and synthetic data. Let L be the Hamming distance and y an attack record, y' a synthetic record, and h a predefined threshold, then a match is considered to have occurred if $\min_{y'} L(y, y') \leq h$.

Then, the $F1$ score can be computed as,

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

Precision and recall are both calculated based on the number of matched records that are in the training set. The $F1$ score is the harmonic mean of precision and recall, representing both equally. $F1$ score indicates the success of attacks. The established threshold stated in a follow up paper for $F1$ score is 0.2 (El Kababji et al., 2023), but there was no indication that this threshold is used across the field.

4. Discussion

4.1. Methodology

While there is certainly no shortage of metrics for assessing disclosure risk in literature, they are not without their issues. A glaring concern with the methods we presented above is the lack of guidance around the threshold used for the suggested metrics, or a lack of consensus across literature around any proposed thresholds. Equation (1) is another interesting thought experiment, it is not useful in a practical situation as is. The authors estimated risk by comparing the risk of releasing the synthetic dataset versus not releasing. Since there is a need to create a synthetic dataset there was never an option to not release the data in the first place. Membership disclosure techniques found themselves to be more based on familiar statistical terms instead of arbitrary scores, which could benefit methodologists, since there are preceding acceptable values for, for example, $F1$ scores (even if it is more applicable for categorical attributes). On that note, there are similar metrics as $F1$ score, such as the F_β (Goutte & Gaussier, 2005) which can be considered as an alternative. Another noteworthy issue is that these metrics cannot be compared to each other because, again, there lacks some translational method between metrics.

Some of the metrics can be used to compare the risk between two released datasets, which is much more intuitive to use. The *CAP* score equation (3) is a good way to compare traditional disclosure control methods, e.g. suppression or rounding, versus using synthetic data. In this way, it can also be calculated with the original and the synthetic dataset, and their values can be compared to see the comparative risk. While there still no standard on how large the difference should be between the two values, comparisons could feel less arbitrary than a singular score derived from other methods.

Another point worth consideration is that most of the literature regard prior knowledge as a key factor in determining disclosure risk, however they do not specify how this information is quantified. Models for creating synthetic datasets are not fit to create outlier records as we see in real life, e.g. Walmart among all grocery stores, because models can only regurgitate existing values or, at best, be perturbed in some way from existing data points (Elliot, 2014; Nowok et al., 2016). Only evaluation techniques such as equation (2) which considers the maximum disclosure among all records can potentially take into consideration of outliers. Equations (3), (4) consider the average risk, which while may be skewed by outliers, underestimate the disclosure risk for those particular records.

4.2. Utility

With all disclosure evaluation of any datasets, the disclosure risk of synthetic data cannot be decoupled with its utility. The broader the expectation of utility of the dataset, the more likely the disclosure risk. Some people even venture to say that measuring risk from a file is measuring its utility. F1 score, as seen in equation (4), is calculated directly from precision and recall (Goutte & Gaussier, 2005), which are two metrics traditionally used to determine the utility of a predictive model, meaning that reducing the risk of a disclosure directly conflicts with maintaining high utility simultaneously. Bayesian techniques, like equation (2), estimates how close a guess can be made based off of prior information. If utility is maximised, i.e. releasing the original data, the score metric calculated will be very high. A synthetic dataset perfectly replicating the original dataset would also result in a high score. However, for confidentiality a somewhat low score is recommended, meaning for Bayesian techniques there is also inherent conflict between confidentiality and utility.

One of the key assumptions that many of the methods have been the potential release of several copies of synthetic datasets from the model. The main advantage of producing multiple datasets is data augmentation, which means any inference or models made from the synthetic datasets are potentially more robust. The risk of disclosure increases as the number of copies increase, so there needs to be good justification for releasing multiple copies.

The release of the synthetic model itself has potentially disclosive consequences. Many of the techniques used to measure disclosure include knowledge of the synthetic model as one of the variables considered, but do not specify how this notion should be quantified. There are obvious benefits in releasing the model or information of the model for its users. The users can create more copies of the synthetic data for more utility, place more trust in the model and the methodology and contribute to more development of synthetic model in the greater schema. However, releasing the model without prudence could potentially lead to high disclosure risk.

4.3 Perceived Risk

Besides the actual risk of disclosure, a more insidious problem is the perceived risk of disclosure. While one can logically conclude that there is no reason this synthetic record contains any real information, the actual respondent whose data looks suspiciously close to the synthetic record may feel that their privacy is at risk. For any statistical organizations that built its foundation on trust, this conclusion is incredibly detrimental. Fully synthetic records might end up matching original ones. This is suspicious but not necessarily problematic. The actual question on that front is whether this is due to a poor implementation of the synthetic file itself (real match) or a result of the randomness of the method (artificial match). The later option is, of course, more desirable.

5. Recommendations

In the light of this literature review, the most actionable method to lowering disclosure risk of synthetic data is to decrease potential knowledge of the real dataset by an attacker by altering the real data prior to synthesis (Sallier, 2021). We recommend the following actions for continuous attributes to limit the extent of attribute disclosure: top/bottom-coding, rounding to an appropriate position for the attribute, and regrouping categories into broader categories.

As highlighted previously, outliers can be problematic; therefore we also recommend limiting the model used in synthesis. This would differ depending on the model used. For decision tree-based models one can set the minimum bucket size for the final node (Gauvin, 2021). Alternatively for outliers, one may apply smoothing if the data isn't overly skewed.

Finally, from our experience and reading it's obvious that when synthetic data is produced at a high level, it can appear very similar to real data. This could cause fear of disclosure in users, even when the records are all synthesized. Therefore, it is crucial for all synthetic data to be well labeled and identified to be synthetic in order to prevent such issues. Synthetic data generation is still an emerging access solution and proper communication around those data products is critical for their optimal use.

6. Conclusion

We reviewed the different types of disclosure risk that affects synthetic data in this paper and some examples of existing disclosure risk assessment techniques. With increased attention of synthetic data as a promising tool for disclosure control, there will be more demand for these techniques to be practical. Existing metrics in literature still require further research to form a standardized procedure for their utilization. We also want to stress that synthetic data is not to replace real data. It is an intermediate step that would help researchers streamline their analysis, since real data access can be inconvenient. Nonetheless, we are hopeful that the maturing of disclosure risk assessment techniques for synthetic data will only support the practical use of synthetic data itself at Statistics Canada.

References

- El Kababji, S., Mitsakakis, N., Fang, X., Beltran-Bless, A., Pond, G. R., Vandermeer, L., Radhakrishnan, D., Mosquera, L., Clemons, M. J., and El Emam, K. (2023), "Can synthetic data accurately mimic oncology clinical trials?", *Journal of Clinical Oncology*, 41(16_suppl), p. 1554. Paper available at https://10.1200/JCO.2023.41.16_suppl.1554.
- Elliot, M. (2014), *Final Report on the Disclosure Risk Associated with the Synthetic Data*, Produced by the SYLLS Team.
- Gauvin, H. (2021), "Generating smart deep files: the example of synthesizing hierarchical data" in *Proceedings of Statistics Canada Symposium 2021 Adopting Data Science in Official Statistics to Meet Society's Emerging Needs*, Catalogue No. 11-522-x.
- Goutte, C., and Gaussier, E. (2005), "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation", In Losada, D.E., Fernández-Luna, J.M. (eds) *Advances in Information*, pp. 345-359. Paper available at https://doi.org/10.1007/978-3-540-31865-1_25.
- Hittmeir, M., Mayer, R., and Ekelhart, A. (2020), "A Baseline for Attribute Disclosure Risk in Synthetic Data", In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, pp. 133–143. Available at <https://dl.acm.org/doi/10.1145/3374664.3375722>

- Hornby, R., and Hu, J. (2021), *Bayesian Estimation of Attribute Disclosure Risks in Synthetic Data with the AttributeRiskCalculation R Package*. Unpublished manuscript. Available at <https://arxiv.org/abs/2103.09805>
- Hu, J. (2021), *Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data*. Ithaca: Cornell University Library, arXiv.org. <https://10.48550/arxiv.1804.02784>. Retrieved from Publicly Available Content Database <https://search.proquest.com/docview/2071976619>
- Hu, J., Reiter, J. P., and Wang, Q. (2014), “Disclosure Risk Evaluation for Fully Synthetic Categorical Data”, *Privacy in Statistical Databases*, pp. 185–199. Springer International Publishing. Available at https://10.1007/978-3-319-11257-2_15.
- Nowok, B., Raab, G. M., and Dibben, C. (2016), *synthpop: Bespoke Creation of Synthetic Data in R*. Foundation for Open Access Statistic. Available at <https://10.18637/jss.v074.i11>.
- Reiter, J. P. (2005), *Estimating Risks of Identification Disclosure in Microdata*. Informa UK Limited. Available at <https://10.1198/016214505000000619>.
- Reiter, J. P., & Mitra, R. (2009). Estimating Risks of Identification Disclosure in Partially Synthetic Data. *The Journal of Privacy and Confidentiality*, 1(1), pp. 99-110. Paper available at <https://10.29012/jpc.v1i1.567>.
- Reiter, J. P., Wang, Q., and Zhang, B. (2014), “Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data”, *The Journal of Privacy and Confidentiality*, 6(1). Paper available at <https://10.29012/jpc.v6i1.635>
- Sallier, K. (2021), “Synthetic Data for National Statistical Organizations: A Starter Guide: Methods and Recommendations”, Paper presented in the Synthetic Data Webinar at the Annual Workshop of the UNECE HLG-MOS.
- UNECE (2022), *Synthetic Data for Official Statistics: A Starter Guide*. United Nations Economic Commissions for Europe. Available at <https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf>.
- Zhang, Z., Yan, C., and Malin, B. A. (2022), “Membership inference attacks against synthetic health data”, *Journal of Biomedical Informatics*, 125. Paper available at <https://10.1016/j.jbi.2021.103977>.
- Zhao, J., Chen, Y., and Zhang, W. (2019), “Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions”, *IEEE Access*, 7, pp. 48901–48911. Paper available at <https://10.1109/ACCESS.2019.2909559>.