SynPrivacy: An Open Framework and Fair Metric for Evaluating Synthetic Data Privacy Risks

Bing Hu[1,2], Asma Bahamyirou[2], Yixin Li[3], Helen Chen[3]
[1]David R. Cheriton School of Computer Science, University of Waterloo
[2]Data Management, Innovation and Analytics, Public Health Agency of Canada
[3]School of Public Health Sciences, University of Waterloo
b25hu@uwaterloo.ca

Abstract

The use of synthetic data in health applications raises privacy concerns, yet the lack of open frameworks for privacy evaluations has slowed its adoption. A major challenge is the absence of accessible benchmark datasets for evaluating privacy risks, due to difficulties in acquiring sensitive data. To address this, we introduce SynPrivacy , an open framework for benchmarking privacy in synthetic data generation (SDG) using simulated sensitive data, ensuring that original data remains confidential. We also highlight the need for privacy metrics that fairly account for the probabilistic nature of machine learning models. As a demonstration, we use SynPrivacy to benchmark CTGAN and propose a new identity disclosure risk metric that offers a more accurate estimation of privacy risks compared to existing approaches. Our work provides a critical tool for improving the transparency and reliability of privacy evaluations, enabling safer use of synthetic data in health-related applications.
Code available at https://github.com/bing1100/simuldata.

1 Introduction

Despite great potential benefit through applied machine learning, access to sensitive personal information requires lengthy approval processes and often with stringent governing rules [1, 2]. Synthetic data is a privacy-enhancing technology (PET) that provides additional protection of sensitive information for data sharing and enables findable, accessible, interoperable, and reusable (FAIR) data standards [3]. Generative artificial intelligence methods such as GAN [4], VAE [4], and DDPMs [5] have shown great potential in generating high-quality synthetic data with high utility and fidelity while enhancing privacy protection when compared to the real data.

Regardless of well defined and established metrics for evaluating the utility and fidelity of synthetic data [6], there is a lack of established open framework and fair metrics for the evaluation of privacy risks, such as re-identification, membership attack and attribute inference attack [7, 8]. Given the primary concern of synthetic data in health applications is privacy, the lack of an open framework for privacy evaluations hinders the adoption of synthetic data technologies [3]. A key challenge for privacy evaluations is the lack of open data for benchmarking of privacy evaluation metrics for synthetic data generation (SDG) due to the difficulty to access identifying data at original source [9, 7, 10]. Open datasets are typically fully de-identified or contain minimal identifiable information, making them unsuitable to benchmark SDG methods for privacy risk evaluation [11]. Conversely, studies that conduct privacy evaluations for their SDG methods often rely on proprietary or private datasets containing identifying information which makes it challenging to compare and benchmark their SDG model against other models [7]. Consequently, SDG publications frequently omit privacy evaluations, despite privacy being a critical aspect, or a main motivation for synthetic data generation [5, 12, 4]. Omitting such key indicators makes it hard for decision-makers for the adoption of synthetic data technology as a viable PET [13, 10].

We introduce SynPrivacy , an open framework designed for benchmarking privacy evaluations of synthetic data using simulated pseudo-identifiable data constructed from non-identifiable real data. This open framework enables AI researchers to benchmark their SDG models in privacy risk evaluation, making SDG models more comprehensive and actionable for policy- and decision-makers. By leveraging SDG as a PET, ad-

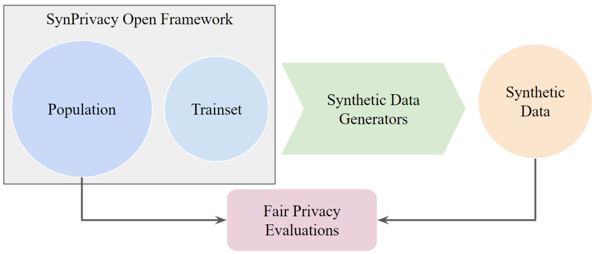

Figure 1: SynPrivacy Open Framework and workflow for synthetic data fair privacy evaluation.

ministrative data holders can safely mobilize their data, thereby accelerating research, collaboration and innovation. This is especially beneficial for AI innovations in healthcare.

Our main contributions are as follows:

- We propose a novel benchmark framework SynPrivacy standardizing the privacy risk evaluation of synthetic data generation models.
- We define a novel concept of fair identity disclosure risk to more accurately evaluate the privacy risks of synthetic data.
- We demonstrate empirical evidence of applying SynPrivacy and fair identity disclosure risk to CT-GAN.

2 Methodology

As shown in Figure 1, the SynPrivacy framework generates a simulated population which a subset is used to train an SDG to generate synthetic data. The generated synthetic data along with the population can then be used for fair privacy evaluations. Through simulated data, the primary aim of the SynPrivacy framework is to provide a privacy benchmark for SDG methods where previously none was available.

2.1 Open SynPrivacy Framework

The idea of SynPrivacy is to simulate identifiable data for real datasets that do not contain identifying data. By simulating quasi-identifiers for de-identified real datasets, SynPrivacy can generate simulated open datasets for which SDGs can be evaluated upon.

SynPrivacy initiates the simulation process by seeding a population with quasi-identifiers following real data distributions. Once the population is established with these simulated quasi-identifiers, non-identifying datasets for real use cases are then linked to each seed quasi-identifier row. Through linking real use case non-identifiable data with seeded quasi-identifier data, we constructed our simulated pseudo-identifiable population and SDG training dataset. Synthetic data from SDG models trained on the simulated data can then be evaluated using our proposed fair privacy risk metrics and other synthetic data evaluation frameworks [14].

2.1.1 Quasi-Identifiers

Quasi-identifiers are pieces of information that are not unique identifiers by themselves, but instead can be correlated together to create a unique identifier [15, 9]. For example, we collect and seed quasi-identifier distributions for age, gender, marital status, occupation, ethnicity, and address.

As the first seed of our simulated population, we apply inverse transform sampling for age using distributions gathered from the census [16]. Age values span between 0 to 99. As gender is correlated with age, we then conditionally sample gender given age for *men+* or *women+*.

For each sampled age and gender, we random sample for marital status, occupation, ethnicity, and address for each row. We apply random sampling on a collected list of 1154 occupations, 7 marital statuses, and 250 ethnicities for each row. Random CA US addresses are generated using an available Python package [17], our framework is not limited to US addresses and other methods for random addresses can be easily applied. Generated random addresses include street and street numbers, city, state, and postal code. For our current work, we do not consider possible correlations between age and gender to occupations, marital statuses, addresses, and ethnicities.

2.1.2 Linked Real Data

We complete our simulated population data by linking non-identifiable real-use case data to our seeded quasi-identifier data. For demonstration we link diabetes data from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) [18] and publicly available BMI data [19]. The diabetes dataset is a non-identifiable dataset with data columns of the number of pregnancies, glucose concentration, diastolic blood pressure, skin thickness, insulin, and BMI, with age as the only quasi-identifier. The BMI data is a non-identifiable dataset with data columns of gender, height, weight, and BMI.

Taking our seed quasi-identifier population, we conditionally sample height, weight, and BMI given gender using the distributions defined in our BMI data for our simulated population. Using BMI and age in our simulated population, we find the nearest neighbour in the diabetes dataset to infill diabetes data columns for our simulated population. Other sampling approaches such as k-nearest neighbours can be applied to ensure fidelity of the infilled rows with minimal degradation from the real data distribution [20].

2.2 Fair Synthetic Data Privacy Evaluations

An open framework for evaluating the privacy of synthetic data necessitates the development of fair and robust privacy metrics that account for the probabilistic nature of machine learning models [3].

De-identified data and synthetic data are generated using fundamentally different methodologies, each resulting dataset with distinctive characteristics. Whereas de-identification predominantly involves logical processes, synthetic data generation is primarily probabilistic. Due to this probabilistic nature of SDG, small variations in numerical data columns are likely to occur between very similar real and synthetic data.

2.2.1 Fair Identity Disclosure Risk (FIDR)

Privacy metrics such as identity disclosure risk (IDR) [7] rely on cardinality of exact matches of numerical identifiers are not necessarily fair measures of the privacy risk of synthetic data generated using probabilistic models. Identity disclosure risk can be simplified to two parts: Real-to-Synthetic Identification Risk, and Synthetic-to-Real identification Risk [7]. The IDR risk is expressed in eq. 1 [7]:

$$IDR = max\left(\frac{1}{N}\sum_{s=1}^{n}\left(\frac{1}{f_s} * I_s\right), \frac{1}{n}\sum_{s=1}^{n}\left(\frac{1}{F_s} * I_s\right)\right) \tag{1}$$

Where N, n is the number of records in the real dataset and synthetic datasets respectively, F_s, f_s is the size of the set of records with the same quasi-identifier values as record **s** in the real data and synthetic data respectively, and I_s is the binary indicator of whether a record **s** in the real data *exactly* matches a record in the synthetic data.

Given the low likelihood of exact numerical matches between very similar real and synthetic data, IDR and other privacy metrics relying on exact matches potentially underestimate the privacy risk of synthetic data. We propose a fair identity disclosure risk (FIDR) that takes into account the variability of SDG models in producing small numerical variations. For FIDR, we propose a new definition for binary match indicator I_s that takes into account small numerical variations. For FIDR, we define binary indicator I_s as 1 if a record **s** in real data and a record SD_j in synthetic data matches exactly on categorical data columns and with cumulative differences less than ϵ on numerical data columns and 0 otherwise.

3 Results

Through SynPrivacy , we are able to simulate a pseudo-identifiable dataset from a de-identified diabetes dataset that can be used to evaluate the privacy risk of various SDG models. We demonstrate SynPrivacy by evaluating IDR and FIDR on synthetic data generated from CTGAN [4] trained from our simulated population. Using SynPrivacy we simulated a population of 3000 rows for the diabetes dataset using our above methodology. From this simulated population, we sample 1000 rows for our training set with quasi-identifier columns age, gender, city, marital status, and data columns BMI, number of pregnancies, glucose, blood pressure, skin thickness, and insulin. We train a CTGAN model with default hyperparameters from the Python library Synthetic Data Vault [21] on our training set to generate 1000 synthetic data rows.

When computing the IDR for the generated synthetic data compared to our population data, the risk score is 0.003. Compared to the IDR, the FIDR with off-by-1 error $\epsilon = 1$ on numerical values of the synthetic data compared to population data is 0.026. Although both risks are below the 0.09 threshold as set by Health Canada and the European Medical Agency [22], we see that IDR considerably underestimates the privacy risk when taking into account small numerical differences between the synthetic and real data. When compared to IDR, FIDR is a much more conservative estimate of risk given the probabilistic nature of SDG methods.

4 Conclusion

Through SynPrivacy , we are able to simulate a pseudo-identifiable dataset from any de-identified dataset that can be openly shared and used to evaluate the privacy risk of various SDG models. We demonstrate SynPrivacy using a diabetes dataset and with CTGAN model. Additionally, we present a fair identity disclosure risk that better considers the probabilistic nature of SDG models. We show that identity disclosure risk can vastly underestimate privacy risk when compared to our fair identity disclosure risk. For future work, we will aim to further develop our SynPrivacy framework, apply our framework to additional models and evaluations, and publish an extensive open dataset for future SDG benchmarking purposes.

References

[1] T. Kokosi, B. L. D. Stavola, R. Mitra, L. Frayling, A. R. Doherty, I. Dove, P. Sonnenberg, and K. L. Harron, "An overview of synthetic administrative data for research," *International Journal of Population Data Science*, vol. 7, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249036803

[2] N. Dattani, P. Hardelid, J. Davey, and R. Gilbert, "Accessing electronic administrative health data for research takes time," *Archives of Disease in Childhood*, vol. 98, pp. 391 – 392, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:1803595

[3] S.-F. Tsao, K. Sharma, H. Noor, A. Forster, and H. H. Chen, "Health synthetic data to enable health learning system and innovation: A scoping review," *Studies in health technology and informatics*, vol. 302, pp. 53–57, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258785381

[4] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Neural Information Processing Systems*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195767064

[5] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "Tabddpm: Modelling tabular data with diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 17 564–17 579.

[6] B. Hu, M. A. Basri, A. Y. M. Abdullah, S.-F. Tsao, Z. Butt, and H. Chen, "Evaluation methods for synthetic data in pursuit of open data," *Journal of Computational Vision and Imaging Systems*, vol. 9, no. 1, pp. 30–33, 2023.

[7] K. El Emam, L. Mosquera, and J. Bass, "Evaluating identity disclosure risk in fully synthetic health data: Model development and validation," *J Med Internet Res*, vol. 22, no. 11, p. e23139, Nov 2020. [Online]. Available: http://www.jmir.org/2020/11/e23139/

[8] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[9] K. El Emam, *Guide to the de-identification of personal health information*. CRC Press, 2013.

[10] D. G. Gomes, P. Pottier, R. Crystal-Ornelas, E. J. Hudgins, V. Foroughirad, L. L. Sánchez-Reyes, R. Turba, P. A. Martinez, D. Moreau, M. G. Bertram *et al.*, "Why don't we share data and code? perceived barriers and benefits to public archiving practices," *Proceedings of the Royal Society B*, vol. 289, no. 1987, p. 20221113, 2022.

[11] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[12] A. Solatorio and O. Dupriez, "Realtabformer: Generating realistic relational and tabular data using transformers," *ArXiv*, vol. abs/2302.02041, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256615552

[13] M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data cards: Purposeful and transparent dataset documentation for responsible ai," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1776–1826.

[14] Y. Yuan, Y. Liu, and L. Cheng, "A multi-faceted evaluation framework for assessing synthetic data generated by large language models," *arXiv preprint arXiv:2404.14445*, 2024.

[15] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[16] S. C. Government of Canada, "Topics, 2021censusage, sex at birth and gender," Feb 2023. [Online]. Available: https://www12.statcan.gc.ca/census-recensement/2021/rt-td/age-eng.cfm

[17] Neosergio, "Neosergio/random-address: This is a provider from a list of real of random addresses that geocode successfully." [Online]. Available: https://github.com/neosergio/random-address

[18] C. F. Turner, H. Pan, G. W. Silk, M.-A. Ardini, V. Bakalov, S. Bryant, S. Cantor, K.-y. Chang, M. DeLatte, P. Eggers *et al.*, "The niddk central repository at 8 years—ambition, revision, use and impact," *Database*, vol. 2011, p. bar043, 2011.

[19] Y. Ersever, "500 person gender-height-weight-body mass index," Jul 2018. [Online]. Available: https://www.kaggle.com/datasets/yersever/500-person-gender-height-weight-bodymassindex

[20] L. E. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[21] N. Patki, R. Wedge, and K. Veeramachaneni, “The synthetic data vault,” in *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 2016, pp. 399–410.

[22] J. Branson, N. Good, J.-W. Chen, W. Monge, C. Probst, and K. El Emam, “Evaluating the re-identification risk of a clinical study report anonymized under ema policy 0070 and health canada regulations,” *Trials*, vol. 21, no. 1, pp. 1–9, 2020.