UNIVERSITY OF WATERLOO

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

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

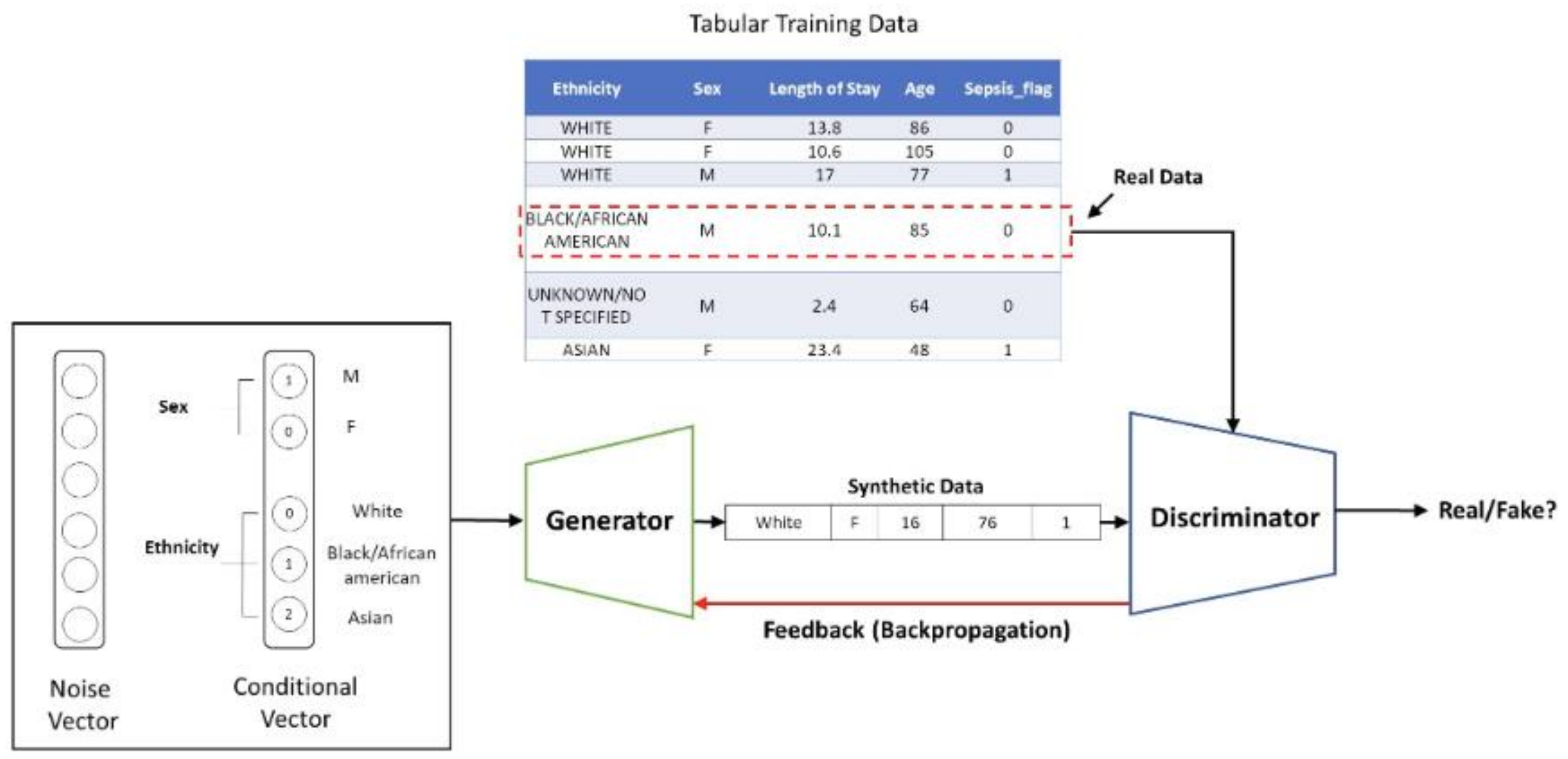
Annual Conference on Vision & Intelligent Systems
CVIS conference

Introduction

Background

Synthetic Health Data

Synthetic health data replicates the structure of real data while original data remains confidential, enabling research and AI model development without patient risks. Generative artificial intelligence methods such as CTGAN have shown great potential in generating high-quality synthetic data with high utility and fidelity while enhancing privacy protection.



Ethnicity	Sex	Length of Stay	Age	Sepsis_flag
WHITE	F	13.8	86	0
WHITE	F	10.6	105	0
WHITE	M	17	77	1
BLACK/AFRICAN AMERICAN	M	10.1	85	0
UNKNOWN/NOT SPECIFIED	M	2.4	64	0
ASIAN	F	23.4	48	1

Lack of Benchmark

Due to difficulty to access identifying data at original source, there is a lack of established open framework and fair metric for the evaluation of privacy risks in synthetic data generation (SDG), such as re-identification, membership attack and attribute inference attack, which hinders the adoption of synthetic

Goal

The objective is to create an open framework, SynPrivacy, providing a privacy benchmark for SDG methods, to improve the accuracy and fairness of privacy metrics and support safer use of synthetic data in health applications.

Methods

Synprivacy Framework

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.

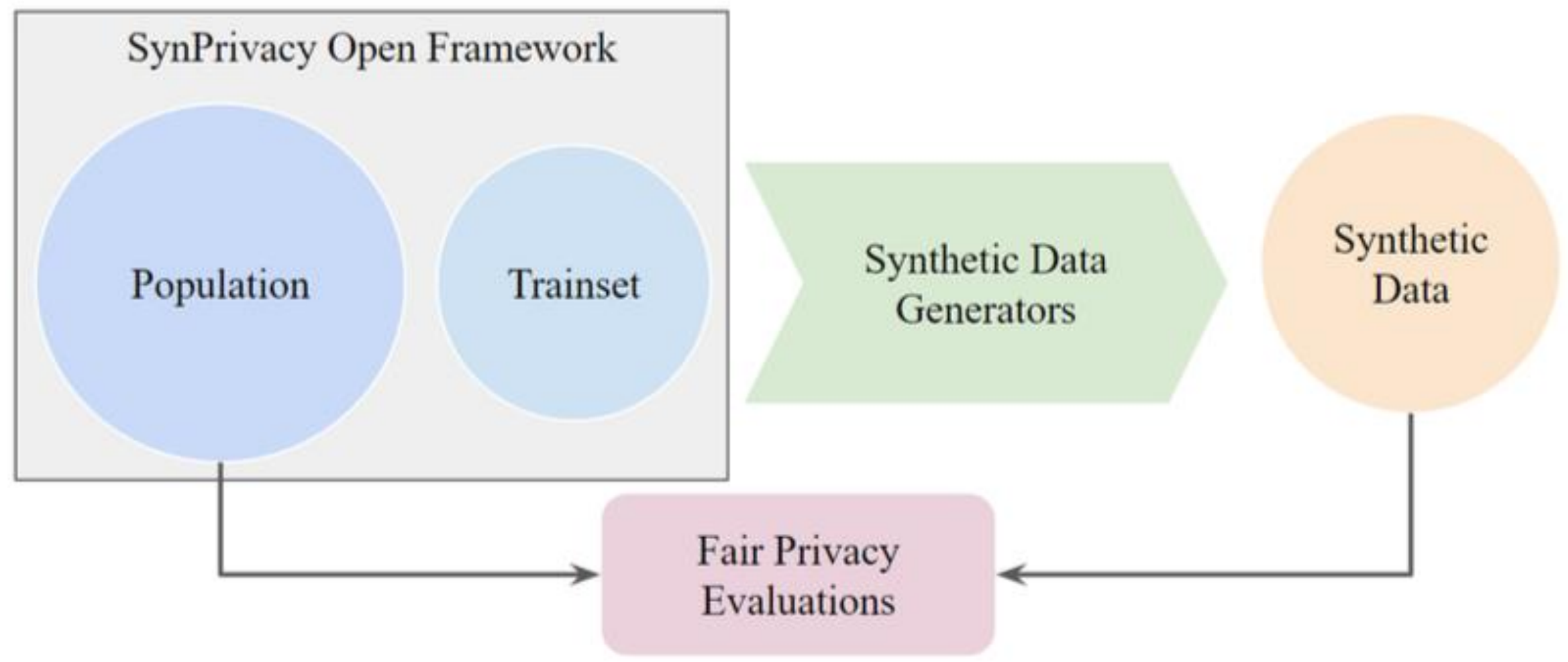


Figure 1. SynPrivacy Open Framework

Quasi-identifiers and Linked Real Data

As shown in Figure 2, Quasi-identifiers are pieces of information that are not unique identifiers by themselves, but instead can be correlated together to create a unique identifier [1,2]. For example, we collect and seed quasi-identifier distributions for age, gender, marital status, occupation, ethnicity, and address.

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 National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) [3] and publicly available BMI data [4]. The diabetes dataset is a non-identifiable dataset with data columns of number of pregnancies, glucose concentration, diastolic blood pressure, skin thickness, insulin, and BMI, with age the only quasi-identifier. The BMI data is a non-identifiable dataset with data columns of gender, height, weight, and BMI.

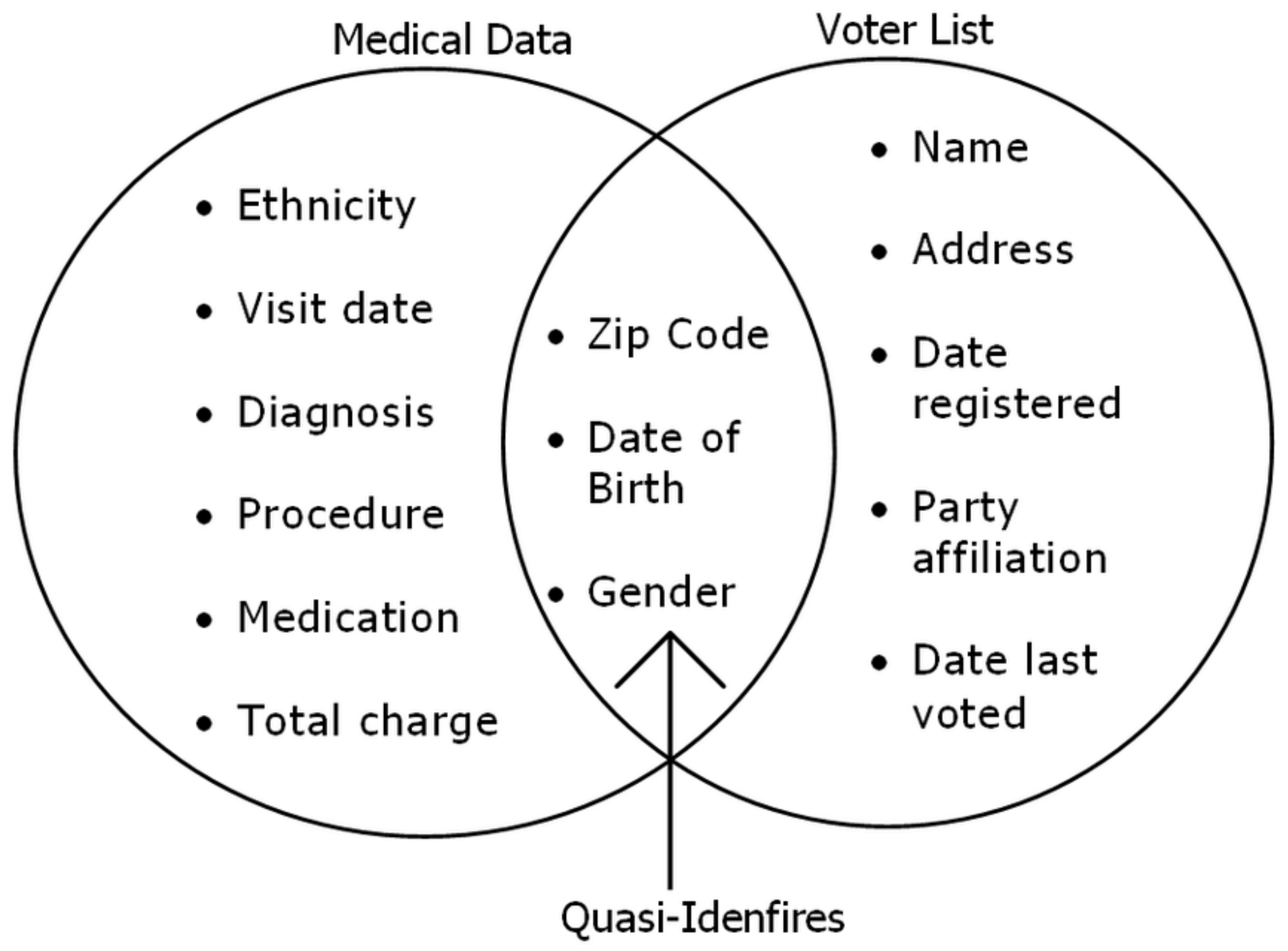


Figure 2. Quasi-identifiers

Fair Identity Disclosure Risk (FIDR)

As shown in Figure 3, privacy metrics such as identity disclosure risk (IDR) [5] 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. 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 new definition for binary match indicator I_s that takes into account small numerical variations.

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

Figure 3. Identity Disclosure Risk (IDR)

Results

The identity disclosure risk (IDR) of the synthetic data was 0.003, while the fair identity disclosure risk (FIDR), with an off-by-1 numerical error (ε = 1), was 0.026. Both values are below the 0.09 threshold set by Health Canada and the European Medical Agency, with FIDR providing a more conservative estimate of privacy risk, reflecting sensitivity to small numerical differences.

Conclusion

Through SynPrivacy, pseudo-identifiable datasets are simulated from de-identified datasets, enabling open sharing and robust evaluation of privacy risks in synthetic data generation (SDG) models.

Demonstrated on a diabetes dataset using the CTGAN model, SynPrivacy introduces a novel FIDR metric, which more accurately reflects the probabilistic nature of SDG models and highlights significant underestimations by traditional identity disclosure metrics.

Future Directions

- Advancing the SynPrivacy framework by optimizing the distribution of quasi-identifiers and applying it to diverse models and evaluations
- Releasing an extensive open dataset to benchmark SDG models, promoting enhanced privacy assurance in synthetic data research.

[1] L. Sweeney, “k-anonymity: A model for protecting privacy,” International journal of uncertainty, fuzzi- ness and knowledge-based systems, vol. 10, no. 05, pp. 557–570, 2002.
[2] K. El Emam, Guide to the de-identification of per- sonal health information. CRC Press, 2013.
[3]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.
[4] Y. Ersever, “500 person gender-height-weight- body mass index,” Jul 2018. [Online]. Avail- able: https://www.kaggle.com/datasets/yersever/ 500-person-gender-height-weight-bodymassindex
[5] 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/