Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220420

Generation and Evaluation of Synthetic Data in a University Hospital Setting

Bayrem KAABACHI[a,b,1], Jérémie DESPRAZ[a] , Thierry MEURERS[c], Fabian PRASSER[c] and Jean Louis RAISARO[a]
[a] *Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland*
[b] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland*
[c] *Berlin Institute of Health @ Charité – Universitätsmedizin Berlin, Germany*

Abstract. In this study, we propose a unified evaluation framework for systematically assessing the utility-privacy trade-off of synthetic data generation (SDG) models. These SDG models are adapted to deal with longitudinal or tabular data stemming from electronic health records (EHR) containing both discrete and numeric features. Our evaluation framework considers different data sharing scenarios and attacker models.

Keywords. Synthetic Data, Privacy, Medical Research, Generative Adversarial Networks

1. Introduction

Synthetic Data Generation (SDG) models, and especially Generative Adversarial Networks (GANs) [1] are innovative and rapidly evolving tools spanning a wide range of fields that can produce new samples from a reference dataset with similar statistical properties. This approach is increasingly considered for the generation of synthetic electronic health records (EHRs) [2]. However, there is a lack of a standardized evaluation of these tools, both in terms of utility and privacy.

Our main contributions are: (i) improvement of the SynTEG framework, [3], state-of-the-art GAN-based SDG model for longitudinal EHR data, (ii) review of the most used utility metrics for both tabular and longitudinal synthetic medical data (iii) adaptation of a privacy-evaluation [4] with regards to hospital-related data sharing use cases and the ability to model various degrees of adversarial background knowledge, and (iv) development of a unified, practical and modular framework that encompasses both utility and privacy metrics for a systematic evaluation of synthetic medical data.

2. Methods

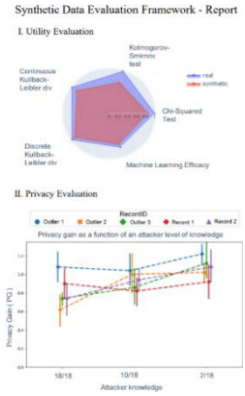
Generative models: We selected the following two state-of-the art GAN-based models: CTGAN [5] for tabular and SynTEG [3] for longitudinal data. We adapted SynTEG to handle longitudinal EHR data that do not contain those diagnosis codes, and that include continuous variables such as vital parameters. We propose a utility evaluation module

[1] Corresponding Author, Bayrem KAABACHI; E-mail: bayrem.kaabachi@gmail.com.

that encompasses the most used utility metrics described in the literature that we organized in six categories. This module characterizes the extent to which our models capture correlations, distributions and temporal patterns of the original dataset. We evaluated the capacity of the models to protect against privacy attacks by adapting and improving the synthetic data privacy evaluation approach developed by Stadler et al. [4]. We enabled the framework to deal with attackers having various partial prior knowledge. We are thus able to measure the risk of membership inference and attribute inference attacks for each patient record. These modules are combined in a unified framework that encompasses both utility and privacy evaluation [2].

3. Results

Results are based on a public dataset provided by the Texas Health Department [6]. The following figure shows a web-based interactive dashboard through which the user can assess the results of multivariate and univariate statistical analysis and visually compare the results between the real and synthetic dataset. We compute the privacy gain PG [4]. A high privacy gain indicates that the probability of re-identification is greatly reduced when disclosing the synthetic data rather than the original dataset. An example report is provided in the Git repository.



4. Discussion and Conclusion

We proposed a unified benchmark to understand the potential gains and risks of using synthetic data. Generative adversarial networks could be a way to fast-forward the development of new AI-based clinical decision support systems or to easily access look-alike data for training and education purposes. Yet, we strongly believe that a formal and common evaluation metric is crucial.

References

[1] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in neural information processing systems; 2014. p. 2672-80.
[2] El Emam K, Mosquera L, Jonker E, Sood H. Evaluating the utility of synthetic COVID-19 case data. JAMIA Open. 2021 Jan;4(ooab012). Available from: https://doi.org/10.1093/jamiaopen/ooab012.
[3] Zhang Z, Yan C, Lasko TA, Sun J, Malin BA. SynTEG: a framework for temporal structured electronic health data simulation. Journal of the American Medical Informatics Association. 2020 Nov;(ocaa262)
[4] Stadler T, Oprisanu B, Troncoso C. Synthetic Data – Anonymisation Groundhog Day. In: 31st USENIX Security Symposium (USENIX Security 22). Boston, MA: USENIX Association; 2022.
[5] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. In: Modeling Tabular Data Using Conditional GAN.Red Hook, NY, USA: Curran Associates Inc.; 2019
[6] TDSHS. Hospital Discharge Data Public Use Data File.

[2] The utility-privacy evaluation framework package can be accessed at https://gitlab.itrcs3-app.intranet.chuv/bkaabachi/evaluation-framework