Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230149

A Framework for Evaluating Synthetic Electronic Health Records

Emmanuella BUDU[a,1], Amira SOLIMAN[a], Kobra ETMINANI[a] and Thorsteinn RÖGNVALDSSON[a]
[a] *Center for Applied Intelligent Systems Research, Halmstad University, Sweden*
ORCiD ID: Emmanuella Budu https://orcid.org/0000-0003-4221-5467

Abstract. Synthetic data generation can be applied to Electronic Health Records (EHRs) to obtain synthetic versions that do not compromise patients' privacy. However, the proliferation of synthetic data generation techniques has led to the introduction of a wide variety of methods for evaluating the quality of generated data. This makes the task of evaluating generated data from different models challenging as there is no consensus on the methods used. Hence the need for standard ways of evaluating the generated data. In addition, the available methods do not assess whether dependencies between different variables are maintained in the synthetic data. Furthermore, synthetic time series EHRs (patient encounters) are not well investigated, as the available methods do not consider the temporality of patient encounters. In this work, we present an overview of evaluation methods and propose an evaluation framework to guide the evaluation of synthetic EHRs.

Keywords. Synthetic data, Electronic Health Records, evaluation

1. Introduction

Synthetic data should maintain the statistical and structural properties of real data without compromising the privacy of the individuals in the real data. Three main criteria: *fidelity*, *utility*, and *privacy*, are used to assess the quality of generated data [1,2]. The *utility* determines the usefulness of synthetic data for predictive and modelling purposes. *Fidelity* assesses the faithfulness of the synthetic data to real data. Finally, *privacy* assesses whether the privacy of the real data is compromised in the synthetic data.

However, several challenges exist. Several comparison methods and measures have been used, and new measures are often introduced in publications. This makes it difficult to compare the data generated by different models as there is no consensus on how to evaluate and compare the synthetic data generated by different models [2]. Secondly, the available methods do not evaluate the variable dependencies or consider the temporality found in patient encounters. They only focus on assessing synthetic EHRs as frozen in time, without dependencies between subsequent entries for the same individuals.

[1] Corresponding Author: Emmanuella Budu, E-mail: emmanuella.budu@hh.se.

2. Methods

In our approach, we first review the available methods and devise a hierarchy of evaluation methods categorized according to the type of data (e.g., categorical, continuous, discrete) and the mode of application (e.g., patient level, cohort level, and feature level). We aim to develop an evaluation framework guided by this hierarchy as illustrated in Figure 1 to assess the quality of synthetic EHRs.

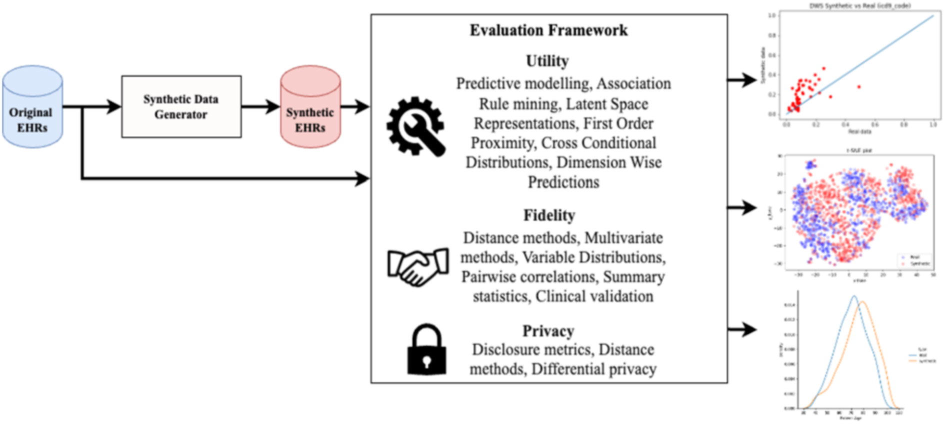
Original EHRs
Synthetic Data Generator
Synthetic EHRs
Evaluation Framework
Utility
Predictive modelling, Association Rule mining, Latent Space Representations, First Order Proximity, Cross Conditional Distributions, Dimension Wise Predictions
Fidelity
Distance methods, Multivariate methods, Variable Distributions, Pairwise correlations, Summary statistics, Clinical validation
Privacy
Disclosure metrics, Distance methods, Differential privacy
DWS Synthetic vs Real (icd9_code)
Real data
Synthetic data
t-SNE plot
Real
Synthetic

Figure 1. Proposed evaluation framework showing the original and synthetic EHRs passed as input to the evaluation framework to obtain qualitative and quantitative assessments. For example, the plots show the distribution of variables in the original and synthetic EHRs.

3. Results, Discussion, and Conclusions

Our categorization of existing evaluation methods identified several different methods of assessment under fidelity, utility and privacy [1,2]. Under fidelity: distance-based methods, variable distribution methods, correlations, comparison of data statistics, clinical validations, and multivariate methods. For utility: predictive modelling, dimension-wise predictions, association rule mining, first-order proximity, latent space representations, and cross-conditional distributions. For privacy: we have disclosure metrics, and distance-based methods.

This work mainly presents an overview of evaluation methods for synthetic EHRs. We propose a framework to standardize the evaluation of synthetic EHRs. We aim to make the implementation of the evaluation framework publicly accessible for use by other researchers in future.

References

[1] Mendelevitch O, Lesh M. Fidelity and Privacy of Synthetic Medical Data, Review of Methods and Experimental Results. ArXiv 2021:39. https://doi.org/10.48550/arXiv.2101.08658.

[2] Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: A systematic review. Neurocomputing 2022;493:28–45. https://doi.org/10.1016/j.neucom.2022.04.053.