AVATAR

Accelerating Innovation through data sharing data with no privacy concerns

Project: CONCEPT/0722/0019

D7, Validation of the synthetic data generator results

Date: 31/01/2024

Contents

1. Introduction

1.1 Scope of the document

The scope of this document is to outline the evaluation results using the synthetic and the available real patient data. The report initially defines the list of evaluation metrics used for validating the generated data samples (whether the generated synthetic data can be distributed easily without revealing any private information) and then the obtained evaluation results using the listed evaluation metrics are presented.

2. Evaluation metrics

2.1 Metrics commonly used in the literature

Metrics commonly used in the literature to test for privacy, resemblance, and utility include the Membership Inference Attack, ground truth tables, receiver operating characteristic (ROC) curve, Hellinger distance, correlations, precision, accuracy and recall [1], [2]. Data statistics (such as the variable distribution, mean and median), correlation differences and the Cox-regression analysis are also used for the evaluation process. Other performance evaluation metrics include the nomogram and the k-fold cross-validation [3].

2.2 Final list of metrics

After the literature review, the final list of metrics was constructed. The final list includes the privacy and utility metrics applicable for the AVATAR project and outlines the metrics that best capture the quality of the artificially generated data. The final list is summarised in Table 1.

Table 1: Performance evaluation metrics.

Metric	Category	Scope	Description
Accuracy	Performance	To evaluate the generator's engine performance	It indicates how well the trained generator represents the statistics of the real data set
Propensity mean squared error (pMSE)	Utility	To define the extent to which the statistical properties of the real data are captured to the synthetic data sets	It predicts whether the synthetic data can be distinguished from the original data
Exact matches	Privacy	To expose how much of the real data may be revealed (directly or indirectly) by the synthetic data	Exact matches (and nearest neighbors distances) between synthetic and original data

3. Performance evaluation results

The results obtained from the application of the artificial intelligence (AI)-powered generator are summarized in Figure 1. The trained AI generator showed high accuracy when representing the statistics of the real patient data, achieving an average accuracy of 96.8%. The high accuracy of the generator indicates it was trained sufficiently and it effectively represents the statistics of the real data set.

Figure 1. Data statistics and utility measures for the real (observed) and synthetic data sets.

The artificially generated synthetic data were then tested for privacy, resemblance, and utility. The obtained results demonstrated pMSE values ranging from 0.06 (PSA before sRT_binarized) to 1.92 (PSA persistence) as shown in Table 2. The low pMSE values obtained indicate that the proposed model can be used to reconstruct accurately the known propensity scores in the synthetic data set. Finally, the exact matches between the real and synthetic data were 6.03% (incidental matches may occur). The generated synthetic data can be thus distributed easily without revealing any private information.

Table 2: Utility metric results.

Characteristic	**pMSE**
Age at sRT	1.594595
pT stage at surgery	1.411748
pT stage at surgery_binarized	1.822839
pT stage at surgery_binarized1	0.136010
R stage	0.289107
R stage_binarized	0.211332
ISUP score in surgery specimen	0.501171
ISUP score in surgery specimen_binarized	0.610310
PSA persistence	1.923894
Local failure - miTr	1.294346
Nodal failure - miN	0.073171
PSA before sRT	0.631925
PSA before sRT_binarized	0.065641

3.1 Discussion

The development and implementation of the AI-powered generator engine for synthetic healthcare data marks a significant milestone in addressing the dual challenges of data scarcity and privacy concerns within the healthcare sector. Through the incorporation of AI generative models and ensemble modelling, this innovative solution offers a promising avenue to generate synthetic data that represent real-world healthcare data while upholding individual privacy. The results demonstrated the effectiveness of the engine in replicating the patient data and underscore its potential to revolutionize healthcare research, innovation, and decision-making processes. By providing access to high-quality, privacy-preserving data sets, the engine empowers researcher and scientists to develop and validate AI-driven healthcare applications/solutions, from predictive analytics to personalized medicine, without compromising sensitive patient information.
Different synthetic versions of the processed data set can be thus generated, maintaining privacy and achieving a high level of resemblance and utility. Those sets can be used to accelerate innovation and enhancing collaboration between public and private. .

4. Conclusions

This document provided the constructed final list with the evaluation metrics. It also presented the evaluation results obtained from the application of the generator using synthetic and real patient data. The obtained results showed the efficacy of developed generator for synthesising fake data without revealing any private information.

5. Acknowledgement

Co-funded by the European Union

RESEARCH & INNOVATION FOUNDATION

Republic of Cyprus

EUNDO
we make it work

German Oncology Center

The project is implemented under the programme of social cohesion “THALIA 2021-2027” co-funded by the European Union, through Research and Innovation Foundation.

References

[1] K. El Emam, L. Mosquera, and R. Hoptroff, *Practical synthetic data generation: balancing privacy and the broad availability of data*. O’Reilly Media, 2020.

[2] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, “Generation and evaluation of synthetic patient data,” *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–41, 2020, doi: 10.1186/s12874-020-00977-1.

[3] C. Zamboglou *et al.*, “Development and Validation of a Multi-institutional Nomogram of Outcomes for PSMA-PET–Based Salvage Radiotherapy for Recurrent Prostate Cancer,” *JAMA Netw. Open*, vol. 6, no. 5, p. e2314748, 2023, doi: 10.1001/jamanetworkopen.2023.14748.