Generation of Realistic Synthetic Validation Healthcare Datasets Using Generative Adversarial Networks

Eda BILICI OZYIGIT [a], Theodoros N. ARVANITIS [a] and George DESPOTOU [a,1]
[a] *Institute of Digital Healthcare, WMG, University of Warwick, UK*

Abstract. *Background:* Assurance of digital health interventions involves, amongst others, clinical validation, which requires large datasets to test the application in realistic clinical scenarios. Development of such datasets is time consuming and challenging in terms of maintaining patient anonymity and consent. *Objective:* The development of synthetic datasets that maintain the statistical properties of the real datasets. *Method:* An artificial neural network based, generative adversarial network was implemented and trained, using numerical and categorical variables, including ICD-9 codes from the MIMIC III dataset, to produce a synthetic dataset. *Results:* The synthetic dataset, exhibits a correlation matrix highly similar to the real dataset, good Jaccard similarity and passing the KS test. *Conclusions:* The proof of concept was successful with the approach being promising for further work.

Keywords. Machine learning, realistic synthetic dataset, privacy, generative adversarial networks

1. Introduction

In recent years, there have been significant advances in digital health that have resulted in applications such as diagnostic, self-monitoring, telehealth and clinical decision support. In addition to their potential benefits, they may also introduce risk to patients, due to technical failures, and unfit clinical validation. Validating them will ultimately need testing the application against datasets, representing the targeted population. However, this a challenging task, as datasets cannot be shared with developers, due to privacy concerns [1]. Even when sharing is possible, their generation and sharing may take several months. The problem is exacerbated by: a) the fact that datasets need to be produced specifically for each application depending on its scope, and b) by the increasing volume of digital health innovations requiring validation. Even if a developer has access to their own datasets, it is difficult for the application to be validated by a third party (e.g. a regulator) without a common dataset. Ultimately, patients may be deprived of potentially beneficial tools. Development of Realistic Synthetic Datasets (RSDs) has been identified as a potential solution to this issue [2, 3], which overcomes privacy concerns, as well as limited utility, of alternatives such as anonymization [1]. These are datasets with completely software generated entries, which exhibit the same statistical

[1] Corresponding Author, G Despotou, Institute of Digital Healthcare, WMG, University of Warwick, Coventry, CV4 7AL, UK; Email: g.despotou@warwick.ac.uk.

properties as (the equivalent) real dataset. Compared with anonymized and de-identified datasets, synthetic datasets have three advantages: a) they overcome the lengthy preparation and approval processes that are still required with anonymized data; b) they offer access to variables that may be considered sensitive and are not included in anonymized and de-identified datasets; and c) they are immune to cross-referencing of information with other datasets (although there are still some concerns to be addressed). Machine Learning (ML) has been a prominent approach to producing large RSDs. ML algorithms are trained based on the real datasets, and then produce a synthetic dataset with similar statistical qualities [4, 5, 6].

The paper presents the results of a Realistic Synthetic Dataset Generation Method (RSDGM) using Generative Adversarial Networks (GANs). GANs use two neural networks; a generator neural network that produces data, and a second, the discriminator, which trains on the real data, classifies the data generated by the former as being true or synthetic, and feeds back its success metric to the generator. The generator network will adapt the data it produces, until the discriminator cannot tell if the generated data are true or false, meaning that the synthetic and the real data are indistinguishable. Figure 1 illustrates how the GAN will begin from random (Gaussian) data (orange points), and end to data similar to the real dataset (blue points).

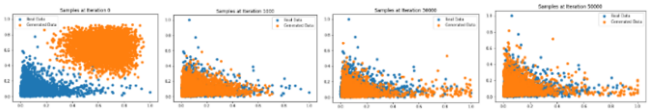

Figure 1. Output during training of the GAN (blue: real data, orange: synthetic data; 50,000 epochs)

The RSDGM was developed as part of a proof of concept project, exploring the feasibility of various methods to generate synthetic datasets, for validation of digital health applications.

2. Method

The RSDGM used the MIMIC III dataset [7] as the real dataset to train on. Figure 2 summarizes the hyper-parameters and data sample used for the experiment.

Hyper-parameters: **Epochs**: 50,000 | **Activation**: ReLu | **Learn. rate**: 1e-4 | **Hidden layers**: 2 | **Loss**: Wasserstein dist. | **Optimiser**: Adam | **Penalty**: 10
Data sample: **No. of patients**: 300 | **No of unique patient admissions**: 379 | **No. of unique lab item & ICD-9 codes**: 1357

Figure 2. Hyper-parameters and data sample

The method covered both numerical and categorical fields from MIMIC III [7], including the ICD-9 diagnosis field, as well as the codes of the lab tests performed during each admission, but was short of producing lab test measurements. A preprocessing stage standardized the values of all variables between 0 and 1. Categorical variables were represented as 0 and 1, using one-hot encoding. A Wasserstein GAN [8] was implemented using two identical neural networks as generator and discriminator using the hyper-parameters shown in figure 2. A series of (informed using the theory) trial and error experiments allowed the selection of the hyper-parameters. Validation of the

synthetic dataset was performed by: a) comparison of correlation matrices of the two datasets, b) Jaccard similarity, and c) the Kolmogorov-Smirnov test. The experiments ran in Anaconda, using Tensor flow, on a GPU enabled machine.

3. Results

The method successfully produced a synthetic dataset based on the selected MIMIC III sample. Figure 3 shows the distribution of a selection of numerical variables of the two datasets, illustrating their similarity. Figure 4 shows an extract (for the same variables) of the correlation matrices of the two datasets.

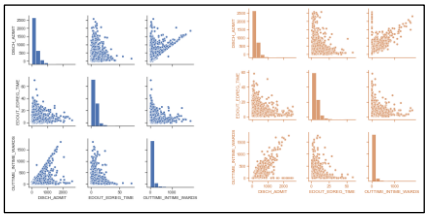

Figure 3. Distribution of a selection of numerical values in the two datasets (blue: real, orange: synthetic)

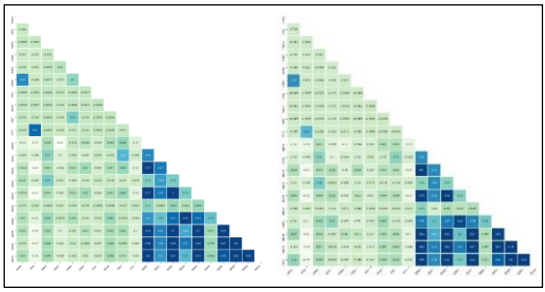

Figure 4. Extract of the two datasets Spearman correlation matrices (left: real, right: synthetic)

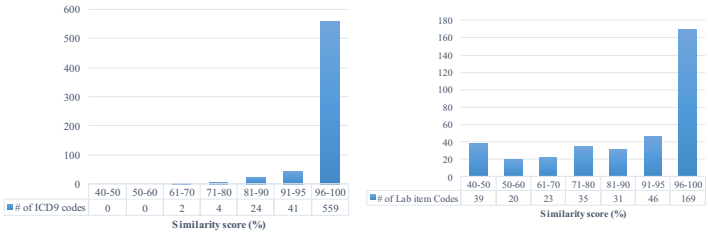

Figure 5. Distribution of Jaccard similarity of ICD-9 and lab test codes

The correlation matrix includes Spearman correlation of all variables including lab tests and ICD-9 codes. Due to the large size of the complete matrix, an extract is presented here. The correlation matrix is a good indication of equivalence of the two datasets, as it can be seen that the association between variables is very similar. Furthermore, not having identical correlations also offers some degree of confidence that the RSDGM has not replicated the datasets, potentially affecting privacy. Figure 5 illustrates the distribution of the Jaccard similarity of the ICD-9 and lab item codes, which tests the distribution similarity of these two types of variables in the two datasets.

Most of the variables (i.e. codes) had very high Jaccard similarity, with some lab codes having low similarity, which is attributed to very low frequency in the training dataset. Finally, a K-S test was performed with $p = 0.05$ to test whether real and synthetic data samples were a subset of the same population, failing to reject the null hypothesis.

4. Conclusions

The developed method was successful in generating a synthetic dataset, using a GAN implementation. Statistical validation of the resultant dataset showed that the two datasets demonstrate very similar statistical qualities. Some observed differences are attributed to low frequency of values and will be explored in the future using bigger samples. Tolerance of the difference between the datasets will need to be justified and accepted, in the context of the validated application. Future work will focus on bigger sample, optimizing the GAN, and justification of statistical tolerance between the datasets. Overall, the GAN based RSDGM showed significant promise.

Acknowledgments

This work was performed in collaboration with the NHS Digital and a2-ci, under the £10m Regulators' Pioneer Fund that the MHRA has been awarded. The fund was launched by The Department for Business, Energy and Industrial Strategy (BEIS) and administered by Innovate UK. The fund enables UK regulators to develop innovation-enabling approaches to emerging technologies and unlock the long-term economic opportunities identified in the government's modern Industrial Strategy.

References

[1] Bellovin SM, Dutta PK, Reitinger N, Privacy and Synthetic Datasets, 22 Stan. Tech. L. Rev. 1 (2019).
[2] Buczak AL, Babin S, Moniz L, Data-driven approach for creating synthetic electronic medical records, BMC Med Inform Decis Mak 10 (2010), 59. https://doi.org/10.1186/1472-6947-10-59
[3] Moniz L, Buczak AL, Hung L, Babin S, Dorko M, Lombardo J, Construction and Validation of Synthetic Electronic Medical Records, Online Journal of Public Health Informatics 1(1) (2009),e2.
[4] Baoqaly MK, Liu C, Chen K, Realistic Data Synthesis Using Enhanced Generative Adversarial Networks, 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE).
[5] McLahlan S, Dube K, Gallagher T, Daley B, Walonoski J, The ATEN Framework for Creating the Realistic Synthetic Electronic Health Record, Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF, pages 220-230.
[6] Schiff S, Gerhke M, Moller R, Efficient Enriching of Synthesized Relational Patient Data with Time Series Data, Procedia Computer Science 141 (2018), 531-538.
[7] Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG, MIMIC-III, a freely accessible critical care database, Scientific Data (2016). DOI: 10.1038/sdata.2016.35.
[8] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A, Improved Training of Wasserstein GANs, https://arxiv.org/pdf/1704.00028.pdf , 2017.