ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

# Quantifying Resemblance of Synthetic Medical Time-Series

Karan Bhanot<sup>1,6</sup>, Saloni Dash<sup>2</sup>, Joe Pedersen<sup>3</sup>, Isabelle Guyon<sup>4</sup> and Kristin P.Bennett<sup>5</sup>

1- Rensselaer Polytechnic Institute - Department of Computer Science - New York - USA
2- BITS Pilani - Department of CSIS - Goa Campus - India

3- Rensselaer Polytechnic Institute - Department of ISE - New York - USA

4- LISN, CNRS/INRIA Université Paris-Saclay - France

5- Rensselaer Polytechnic Institute - Department of Mathematics - New York - USA

6- OptumLabs Visiting Fellow

**Abstract**. Access to medical data is often restricted due to privacy laws e.g. HIPAA and GDPR. We address the viability of substituting real data with synthetic data to protect privacy while maintaining utility. Medical data records are fundamentally longitudinal, with one patient having multiple health events influenced by covariates like gender, age etc. Synthesis of medical data, hence, falls under time-series generative modeling. We demonstrate methods to measure synthetic medical time-series quality on datasets from previously published synthetic data research. We deploy four time-series metrics to quantify resemblance in synthetic and real covariate plots while comparing baseline data generation methods.

# 1 Introduction

Dissemination of medical records is essential to facilitate new research and make effective decisions. For instance, Electronic Medical Records (EMRs) can enhance patient care, facilitate identification of eligible patients, reduce costs and improve efficiency in clinical research [1]. However, access to medical data is often restricted by privacy laws such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR)[2].

Synthetic medical data is an elegant solution that can enable dissemination of high utility data that resembles real data while preserving patient privacy. Medical records are often longitudinal, with each patient having multiple medical events influenced by covariates like age, gender etc. [3] and thus, it is essential that synthetic data generation methods capture the temporal trend in the data. Published research often compares the real and synthetic health data using medical figures such as covariate plots. Dash et. al [3] and Yale et. al [4] both reproduce published figures to illustrate the "utility" of the generated data. For example, Dash et. al plot average minutes slept for each hour of the day for different patient populations identified by the covariate age, and then visually compared the plots as an indicator of resemblance of the real and synthetic data. However, this analysis is qualitative in nature and is subject to user bias and domain expertise. Thus, in this paper, we propose the use of four time-series metrics such as Root Mean Square Error (RMSE), Directional Symmetry (DS), etc. to empirically quantify the real and synthetic medical time-series quality by measuring the resemblance, and discuss the importance of evaluation stratified by covariates in synthetic medical time-series data evaluation.

# 2 Method

# 2.1 Synthetic Data Generation

We generate synthetic medical data using four data generation methods and then compare them with the real data from which they were generated.

The first method uses bootstrapping to sample with replacement from the real training data with the same number of records as in the real data. As the data is bootstrapped from real data records, the covariate plot would resemble real data quite significantly. While this provides an excellent baseline model for comparison since the univariate and multivariate statistics (e.g. the mean and variance of each variable with and without covariate stratification) are preserved, the data cannot be released since it does not preserve patient privacy.

The second and third methods involve randomly permuting the values of each column by 10% and 90% respectively, resulting in two new synthetic datasets. These two datasets have the exact same univariate statistics as the original dataset. While the 10% randomly permuted data would not be much different from the real data, the 90% randomly permuted data might lose the underlying patterns from the real data since join dependencies between variables are severely attenuated. For example, if we stratify data using one of the covariates, the data trends may be lost. However, both act as good baseline models at the two extremes of the spectrum.

Lastly, we use HealthGAN model for generating synthetic medical records as the fourth method. The HealthGAN model handles categorical, continuous, and time-series data as shown by [3] and [4] and is shown to preserve privacy while retaining utility. In contrast to the other three methods, the HealthGAN model attempts to capture the multivariate distribution of the data at the potential expense of capturing the marginal or univariate distribution of each attribute.

#### 2.2 Time-Series Evaluation Metrics

We use time-series evaluation metrics for empirically analysing the resemblance of synthetic data covariate plots. We use metrics from Financial Time-Series Clustering literature [5] as well as Time-Series Clustering [6] to quantify different aspects captured by the synthetic data. Let  $\{x^k\}_{k=1}^n$  represent the real timeseries and  $\{y^k\}_{k=1}^n$  represent the synthetic time-series in the covariate plots.

Root Mean Square Error (RMSE): It measures the closeness of values at every time point between the real and synthetic data. It is calculated as:

$$\sqrt{\frac{1}{n} \sum_{k=1}^{n} (x^k - y^k)^2}$$
(1)

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

We modify the RMSE values such that 1 indicates a perfect match while 0 indicates a complete mismatch, using the function:

$$f(x) = \frac{1}{1+x} \tag{2}$$

**Pearson's Correlation Coefficient (cc):** It quantifies how well the linear correlations present in the real data are captured by the synthetic data. We modify the correlation formula such that 1 implies perfect correlation and 0 implies perfect anti-correlation using the mathematical function:

$$\frac{\frac{\operatorname{cov}(\mathbf{x},\mathbf{y})}{\sigma_x \sigma_y} + 1}{2} \tag{3}$$

**Directional Symmetry (DS):** It is used to measure how well the trend direction (upwards or downwards) in the real data is captured by the synthetic data and is measured using:

$$\frac{1}{n-1} \sum_{k=1}^{n-1} d_k, \text{ where } d_k = \begin{cases} 1, & \text{if}(x^{k+1} - x^k)(y^{k+1} - y^k) \ge 0\\ 0, & \text{otherwise} \end{cases}$$
(4)

DS with value 1 indicates perfect directional match while 0 indicates otherwise. Short Time-Series Distance (STS):

It measures the distance between the finite difference approximation of the derivatives of the two time series. We calculate STS as:

$$\sqrt{\sum_{k=1}^{n-1} \left(\frac{y^{k+1} - y^k}{t^{k+1} - t^k} - \frac{x^{k+1} - x^k}{t^{k+1} - t^k}\right)^2} \tag{5}$$

We normalize the STS values to the range 0 (bad) to 1 (good) using the formula defined in (2). Note: the difference  $t^{k+1} - t^k$  is always 1 in our study.

These metrics collectively assess the ability of generative methods to produce high quality synthetic time-series data at the covariate level and empirically identify the effectiveness of the metrics in analysing time-series figures.

#### 3 Experimental Results

We use American Time Use Survey (ATUS) data from Dash et. al [3, 7] and medical claims Autism Spectrum Disorder (ASD) data from Yale et. al [8] as case studies for covariate time-series plot evaluation. The synthetic data produced by the HealthGAN model in the papers has been shown to preserve privacy as well as resemble the real data but no quantitative analysis for the time-series has been shown. Thus, we deploy the four time-series metrics to quantify the resemblance of the real and synthetic time-series.

**ATUS data:** ATUS data is a national, annual survey of 30K people in the United States on how they organize their time across various activities. We plot

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.



Fig. 1: Average sleep times for various age groups across the seven days of the week. Synthetic data sets, generated from ATUS by bootstrapped, 10% randomly permuted and HealthGAN, show trends similar to the real data while the 90% randomly permuted data does not capture any temporal trends.

the average sleep times for various age groups across the seven days of the week, as shown in Figure 1, and compare them with the complete real data. Just by looking at the plots, it is difficult to conclude which method is performing the best, suggesting the need for quantitative quality analysis for better empirical comparison. The results of the time-series metrics are shown in Figure 2.

**ASD data:** For our second case study, we use the medical claims Autism Spectrum Disorder (ASD) dataset which includes diagnosis of more than 280,000 children across seven Comorbid Medical Conditions (CMCs). We plot the Sleep disorder prevalence (binary data i.e. presence or absence of diagnosis) across the 5 year enrollment period for the Race covariate. The synthetic covariate plots are compared against the covariate plot for the complete real data and the time-series metrics results are shown in Figure 2. The dataset was accessed inside a secure environment provided by OptumLabs<sup>®</sup>. The several synthetic data generation methods were deployed inside the environment for analysis. The real data cannot be exported outside this secure environment.

For the ATUS data, bootstrap and 10% randomly permuted data have the best results as they closely resemble the real data, demonstrating the effectiveness of these metrics in quantifying quality of synthetic medical time-series. Using the metrics, we are able to empirically conclude that HealthGAN captures the trend but not as effectively as the previous methods. This indicates that privacy preservation comes at a cost of sacrificing covariate time-series quality. ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.



Fig. 2: Synthetic ATUS and ASD results using four different metrics: (a) ATUS data: The 90% randomly permuted data performs the worst amongst all methods with all methods performing bad on Short Time-Series Distance metric; (b) ASD data: All methods are comparable but the HealthGAN performs worst on the Directional Symmetry metric.

The 90% randomly permuted data loses the inherent meaning of patterns and utility, and performs the worst on all metrics. The permutation and bootstrap methods perfectly capture the univariate statistics by design, but the proposed metrics show that there is considerable difference in their overall performance.

For the ASD data, the bootstrap and 10% randomly permuted data perform well on all metrics. The HealthGAN model performs well on RMSE, Pearson, and STS. However, it lags behind the other three models in capturing the directional aspect hinting at the need for further analysis. We also notice that the 90% randomly permuted data performs well. The ASD Sleep data is sparse and binary, thus, permuting the columns does not lead to much deviation from the original data across sub-populations of covariate plots. This is clearly evident in Figure 3, where real data plot is shown against HealthGAN and 90% randomly permuted data covariate plots.



Fig. 3: Prevalence plots for Sleep disorder where 90% randomly permuted data captures the general trend but the HealthGAN shows some deviation across all sub-populations.

We observe that the 90% permutation closely resembles real data implying that permuting binary prevalence Sleep data does not affect the covariate level grouping when the general trend across covariate categories is the same. This explains the high performance of 90% random permutation on the four timeseries metrics. On the other hand, the HealthGAN synthetic data plot shows that there are deviations across all sub-populations of the Race covariate, resulting in lower metric values and suggesting that the model is actually introducing temporal trends at the covariate level that do not exist in the original data.

Baseline models (bootstrapping and random permutation), that perform well on univariate analysis, are not always the best synthetic data generation methods as they might either not preserve privacy or may produce unusable data. The comparison underscores the relevance of using additional covariate-level timeseries metrics to identify the quality of synthetic medical time-series data.

# 4 Discussion and Conclusion

The time-series evaluation metrics used enable us to quantify the quality between real and synthetic medical time-series. The results from the case studies underscore the importance of using these metrics to evaluate resemblance in covariate temporal plots. While the generated synthetic data might resemble the real data overall and preserve privacy, the same might not be true at the granular covariate level. Hence, it is vital that evaluation of synthetic data capture trends across covariates as well using these and potentially other metrics. These methods can also be used to quantify the difference between figures for real and synthetic when assessing utility.

## References

- Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9, 2017.
- [2] European Parliament and of the Council (2016, Apr. 27). Directive 95/46/ec (general data protection regulation). L119, pages 1–88.
- [3] Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, Andrew Yale, and Kristin P Bennett. Synthetic event time series health data generation. arXiv preprint arXiv:1911.06411, 2019.
- [4] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Privacy preserving synthetic health data. In Proceedings of the 27. European Symposium on Artificial Neural Networks ESANN, pages 465–470, 2019.
- [5] Gourav Kumar, Sanjeev Jain, and Uday Pratap Singh. Stock market forecasting using computational intelligence: A survey. Archives of Computational Methods in Engineering, pages 1–33, 2020.
- [6] T Warren Liao. Clustering of time series dataâa survey. Pattern recognition, 38(11):1857– 1874, 2005.
- [7] Saloni Dash, Andrew Yale, Isabelle Guyon, and Kristin Bennett. Medical time-series data generation using generative adversarial networks. In Press at Proceedings of the International Conference on Artificial Intelligence in Medicine AIME 2020.
- [8] Andrew Yale, Saloni Dash, Karan Bhanot, Isabelle Guyon, John S. Erickson, and Kristin Bennett. Synthesizing quality open data assets from private health research studies. In Press at BIS 2020 workshop post-conference proceedings.