# Generating Private and Fair Long-Sequenced Longitudinal Healthcare Records

Md Fahim Sikder[1,a] , Resmi Ramachandranpillai[1] and Fredrik Heintz[1]

***Abstract*— Generating long-sequenced longitudinal healthcare records is critical as it has numerous potential applications. Long-sequenced longitudinal data allow us to better understand and find patterns from the data. However, privacy concerns make it challenging to share the data, and real-world data is not bias-free. Generative Adversarial Networks (GAN) have been used to synthesize healthcare records, but the high dimensionality of these data makes them challenging to generate. From these motivations, we are working on a diffusion-based model that is capable of generating long-sequenced, fair, and private healthcare records.**

## I. Introduction

Generating healthcare data is essential due to its potential applications. For better healthcare diagnosis, research into various diseases, and the development of artificial intelligence-based models to aid automatic analysis, vast amounts of real world data are needed. Unfortunately, these data might contain sensitive patient or test subjects' information. This makes it challenging to use the original data. Even if the data is de-identified, a data breach is possible.

To mitigate this issue, synthetic data has been used widely [1], [2], [3]. Researchers train a generative model to generate synthetic data statistically similar to the original data, which should serve its purposes. These generative models may, however, retain training data and leak it.

Differential privacy [4] is a technique that mathematically guarantees the privacy protection of sensitive data. It protects the data from privacy attacks such as linkage and reconstruction attacks. So, researchers have used a combination of generative models with differential privacy to generate private synthetic data that protect and overcome the data remembering problem of generative models [5].

Healthcare longitudinal records refer to the data collected over time, such as laboratory test results, medication history, blood pressure, etc. Researchers use these longitudinal data to find patterns and make decisions. Longer sequence lengths of data can capture more information about the context of the data.

Again, real-world datasets are biased regarding key demographic factors like race and gender. This kind of bias can be difficult to correct due to the lack of representation of certain groups in the data. Furthermore, these can lead to inaccurate results and conclusions that are not representative of the full population [6], [7], [8], [9].

So, it is essential to have a generative model capable of generating long-sequenced private and bias-free or fair data.

## II. Approach

This project is part of an ongoing PhD study. In the PhD study, we aim to propose methods for generating high-quality and long-sequenced time series data and maintaining privacy and fairness in the model. Also, we aim to propose methods to evaluate the quality of synthetic long-sequence time-series data. So, we are trying to address the following research questions:

1) How to generate long-sequenced time series data?
2) How to make the data/model privacy-preserved?
3) How to make the data/model fair?
4) How to evaluate the data/model with respect to privacy, fairness and fidelity?

We already worked with research problems 1 and 4 (partially). We proposed two frameworks based on Generative Adversarial Networks (GAN) and the Diffusion model. We did extensive experiments on both simulated and real-world data. Both of our frameworks could generate high-quality, long-sequenced time series data. Additionally, we proposed two metrics for evaluating the predictiveness and distinguishability of the synthetic data. This project aims to work with research problems 2 and 3.

Generating longitudinal healthcare records poses a unique challenge as the data itself is high-dimensional, and the feature values change over time. Convolutional Neural Networks or Recurrent Neural Network-based Generative Adversarial Networks (GANs), Variational Autoencoders have been used [1], [2], [3], [5], [10], [11] to synthesize healthcare data. On top of GAN, differential privacy has been used to address the privacy issue in synthetic data [5], [12]. However, GANs are difficult to train and often suffer from mode collapse. CNN and RNN-based architecture are unsuitable for generating long-sequenced data of their limitations in architecture [13], [14].

One approach to mitigate this problem is using diffusion-based generative models [15], [16], and transformers-based architecture [17] in the diffusion process. Diffusion-based models learn the data's semantic nature, which is why they can overcome the mode-collapse problem. Also, the attention mechanism in the transformers architecture allows us to capture long-term dependencies. To address the privacy issue, we are training the generative model in a DP-SGD manner [12]. To address the fairness issue, we are proposing a fairness penalty and adding it to the diffusion training process. Experiments are underway.

For this study, we use MIMIC-III [18] and MIMIC-IV [19] datasets and use *Gender* as the sensitive attribute.

[1] Department of Computer and Information Science (IDA), Linköping University, Sweden

[a] Contact: md.fahim.sikder@liu.se

# References

[1] M. K. Baowaly, C.-C. Lin, C.-L. Liu, and K.-T. Chen, "Synthesizing electronic health records using improved generative adversarial networks," *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 228–241, 2019.

[2] A. H. Z. Nik, M. A. Riegler, P. Halvorsen, and A. M. Storås, "Generation of synthetic tabular healthcare data using generative adversarial networks," in *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*, pp. 434–446, Springer, 2023.

[3] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "Medgan: Medical image translation using gans," *Computerized medical imaging and graphics*, vol. 79, p. 101684, 2020.

[4] C. Dwork, "Differential privacy," in *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pp. 1–12, Springer, 2006.

[5] A. Torfi, E. A. Fox, and C. K. Reddy, "Differentially private synthetic medical data generation using convolutional gans," *Information Sciences*, vol. 586, pp. 485–500, 2022.

[6] A. Rajabi and O. O. Garibay, "Tabfairgan: Fair tabular data generation with generative adversarial networks," *Machine Learning and Knowledge Extraction*, vol. 4, no. 2, pp. 488–501, 2022.

[7] X. Wu, D. Xu, S. Yuan, and L. Zhang, "Fair data generation and machine learning through generative adversarial networks," in *Generative Adversarial Learning: Architectures and Applications*, pp. 31–55, Springer, 2022.

[8] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan+: Achieving fair data generation and classification through generative adversarial nets," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1401–1406, IEEE, 2019.

[9] J. Liu, Z. Li, Y. Yao, F. Xu, X. Ma, M. Xu, and H. Tong, "Fair representation learning: An alternative to mutual information," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1088–1097, 2022.

[10] S. Biswal, S. Ghosh, J. Duke, B. Malin, W. Stewart, C. Xiao, and J. Sun, "Eva: Generating longitudinal electronic health records using conditional variational autoencoders," in *Machine Learning for Healthcare Conference*, pp. 260–282, PMLR, 2021.

[11] L. Mosquera, K. El Emam, L. Ding, V. Sharma, X. H. Zhang, S. E. Kababji, C. Carvalho, B. Hamilton, D. Palfrey, L. Kong, *et al.*, "A method for generating synthetic longitudinal health data," *BMC Medical Research Methodology*, vol. 23, no. 1, pp. 1–21, 2023.

[12] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[13] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep Learning for Time Series Classification: A Review," *Data mining and knowledge discovery*, vol. 33, no. 4, pp. 917–963, 2019.

[14] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 3285–3292, IEEE, 2019.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[19] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, B. Moody, B. Gow, L.-w. H. Lehman, *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.