Advancing Medical Multimodal Learning and Data Generation with Diffusion Model and LLM

Yuan Zhong

The Pennsylvania State University
yfz5556@psu.edu

Abstract

Synthesizing electronic health records (EHR) is essential for addressing data scarcity, bias, and fairness in healthcare models. EHR data are inherently multimodal and sequential, encompassing structured codes, clinical notes, medical images, and irregular time intervals. Traditional generative models like GANs and VAEs struggle to capture these complexities, while diffusion-based models offer improvements but remain limited to task-specific applications. To address these challenges, two diffusion-based models, MedDiffusion and EHRPD, have been developed. MedDiffusion enhances health risk prediction by generating synthetic patient data and capturing visit-level relationships, while EHRPD generates sequential, multimodal EHR data, incorporating temporal interval estimation to improve diversity and fidelity. Future work aims to overcome limitations in multimodal data generation by developing a generalized model capable of handling diverse modalities simultaneously, expanding the applicability of EHR data generation across healthcare tasks.

Introduction

Synthesizing electronic health records (EHR) has become a crucial approach for overcoming challenges related to data insufficiency, bias, and fairness in healthcare predictive models. EHR data are inherently multimodal and sequential, encompassing structured codes (e.g., diagnosis, medication, and procedure codes), unstructured data (e.g., clinical notes), and irregular time intervals between visits. Generating synthetic data that accurately reflects this complexity is essential for building robust healthcare applications and improving model performance.

Various generative modeling techniques have been proposed to tackle this task. Early methods focused on generative adversarial networks (GANs), such as MedGAN (Armanious et al. 2020) and ehrGAN (Che et al. 2017), which aimed to generate patient data by learning from aggregated visit-level representations. However, these models faced limitations in capturing temporal dependencies between visits, a key factor in modeling patient health progression.

To address these limitations, variational autoencoders (VAEs) such as EVA (Biswal et al. 2021) and TWIN (Das, Wang, and Sun 2023) introduced latent variable models for learning the structure of EHR data. While VAEs improved the generation of visit-level details, they relied on simple mapping functions that often overlooked the intricate temporal relationships in EHR data. On the other hand, language models like PromptEHR (Wang and Sun 2022) and HALO (Theodorou, Xiao, and Sun 2023) introduced the use of autoregressive and masked language modeling techniques for EHR generation, though these models often compromised data diversity for quality, limiting their ability to generate nuanced healthcare data.

More recently, diffusion-based models have emerged as a promising solution for medical data generation. TabDDPM (Kotelnikov et al. 2023), Meddiff (He et al. 2023), and ScoEHR (Naseer et al. 2023) represent early diffusion models that improve the quality and diversity of synthetic healthcare data by leveraging noise-based generation processes. These models have demonstrated success in generating high-fidelity data across both categorical and numerical forms, yet they remain focused on task-specific applications and often fail to capture the full temporal and multimodal characteristics of EHR data.

To further advance the field, diffusion-based models like MedDiffusion (Zhong et al. 2024a) and EHRPD (Zhong et al. 2024b) were introduced. MedDiffusion augments synthetic patient data during training, using a step-wise attention mechanism to capture hidden relationships between visits and improve predictive performance in health risk tasks. EHRPD, on the other hand, expands the scope of EHR generation by incorporating temporal interval estimation into the diffusion process, ensuring that generated visit sequences accurately reflect both temporal dependencies and multimodal relationships. These methods have shown that diffusion-based approaches can successfully model the complexity of EHR data while generating realistic and diverse patient records.

However, despite the progress made by diffusion models, a critical challenge remains: the generation of multimodal data is often constrained to one-to-one or two-to-one modality combinations. Moreover, models developed for general domains struggle to handle the nuanced relationships found in medical data, where the alignment between modalities is more subtle and difficult to capture.

Thus, there remains a significant need for advanced multimodal EHR generation methods that can capture the tem-

poral dynamics of patient data and generate across diverse modalities, including EHR codes, clinical notes, medical images, and continuous sensor readings. Addressing this challenge will further enhance the applicability of synthetic EHR data for a wide range of healthcare tasks.

Previous Work

Medical data generation addresses data scarcity, improves quality, and promotes fairness in healthcare models. However, existing methods often suffer from task-unrelated designs that limit their effectiveness in handling complex healthcare data.

MedDiffusion, a diffusion-based model, enhances health risk prediction by generating synthetic patient data during training, expanding the sample space, and improving predictive performance. Using a step-wise attention mechanism, it uncovers hidden relationships between patient visits, significantly outperforming 14 baselines across multiple datasets.

To further address EHR data generation, EHRPD was introduced. This model predicts the next patient visit while incorporating time interval estimations, improving both the quality and diversity of the generated data. Through a time-aware visit embedding module and a predictive denoising diffusion probabilistic model (P-DDPM), EHRPD significantly outperforms existing approaches, advancing the field in terms of fidelity, privacy, and utility.

Future Work

Overcoming Limitations in Multimodal EHR Data Generation

While significant progress has been made in EHR data generation, current methods are constrained by their focus on one-to-one or two-to-one modality generation. This limited scope reduces the applicability of these models, preventing them from fully leveraging the rich diversity of data in EHR systems, including structured codes, unstructured medical notes, medical images, and continuous sensor readings.

In the general domain, a few models c(Zhao et al. 2024; Xu et al. 2023) have successfully achieved simultaneous generation across different modalities. However, these models are tailored for general-purpose tasks and are difficult to directly implement for medical data generation. The primary challenge lies in the nuanced relationships in medical data, where the match between modalities—such as EHR codes and clinical notes—is more complex and subtle compared to general domain tasks.

To address these gaps, my future work focuses on developing a method for generating multimodal EHR data that optimally captures and integrates EHR codes, medical notes, medical images, and continuous readings. By creating a generalized model capable of handling these varied modalities simultaneously, this approach aims to broaden the scope of EHR data generation, enabling more accurate and comprehensive synthetic data that can better serve healthcare applications.

References

Armanious, K.; Jiang, C.; Fischer, M.; Küstner, T.; Hepp, T.; Nikolaou, K.; Gatidis, S.; and Yang, B. 2020. MedGAN: Medical image translation using GANs. *Computerized medical imaging and graphics*, 79: 101684.

Biswal, S.; Ghosh, S.; Duke, J.; Malin, B.; Stewart, W.; Xiao, C.; and Sun, J. 2021. EVA: Generating longitudinal electronic health records using conditional variational autoencoders. In *Machine Learning for Healthcare Conference*, 260–282.

Che, Z.; Cheng, Y.; Zhai, S.; Sun, Z.; and Liu, Y. 2017. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *ICDM*, 787–792. IEEE.

Das, T.; Wang, Z.; and Sun, J. 2023. Twin: Personalized clinical trial digital twin generation. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 402–413.

He, H.; Zhao, S.; Xi, Y.; and Ho, J. C. 2023. MedDiff: Generating electronic health records using accelerated denoising diffusion model. *arXiv preprint arXiv:2302.04355*.

Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; and Babenko, A. 2023. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, 17564–17579.

Naseer, A. A.; Walker, B.; Landon, C.; Ambrosy, A.; Fudim, M.; Wysham, N.; Toro, B.; Swaminathan, S.; and Lyons, T. 2023. ScoEHR: Generating Synthetic Electronic Health Records using Continuous-time Diffusion Models. In *Machine Learning for Healthcare Conference*, 489–508. PMLR.

Theodorou, B.; Xiao, C.; and Sun, J. 2023. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature communications*, 5305.

Wang, Z.; and Sun, J. 2022. PromptEHR: Conditional Electronic Healthcare Records Generation with Prompt Learning. In *Conference on Empirical Methods in Natural Language Processing*.

Xu, X.; Wang, Z.; Zhang, G.; Wang, K.; and Shi, H. 2023. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7754–7765.

Zhao, X.; Liu, B.; Liu, Q.; Shi, G.; and Wu, X.-M. 2024. EasyGen: Easing Multimodal Generation with BiDiffuser and LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1351–1370.

Zhong, Y.; Cui, S.; Wang, J.; Wang, X.; Yin, Z.; Wang, Y.; Xiao, H.; Huai, M.; Wang, T.; and Ma, F. 2024a. MedDiffusion: Boosting Health Risk Prediction via Diffusion-based Data Augmentation. In *SIAM International Conference on Data Mining*.

Zhong, Y.; Wang, X.; Wang, J.; Zhang, X.; Wang, Y.; Huai, M.; Xiao, C.; and Ma, F. 2024b. Synthesizing Multimodal Electronic Health Records via Predictive Diffusion Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4607–4618.