

생성모델에 따른 전자의무기록 임상적 효용성 비교

박진규^{1*}, 김만수¹

광주과학기술원 AI대학원¹

Comparative Clinical Efficacy of Electronic Health Records Based on Generative Models

Jingyu Park¹, Mansu Kim^{1*}

AI Graduate School, Gwangju Institute of Science and Technology, Gwangju, Korea

*airrocket@gm.gist.ac.kr

Abstract

Electronic Health Records (EHRs) are crucial for research. However, they contain real patient details and there are privacy concerns. Even with de-identified data, there's a risk of patient re-identification through malicious actions. To solve this problem, research generating synthetic data which is similar to actual data but without real patient information is important. This paper introduces a model that can process both time-invariant and time-dependent variables in EHR data. We train a random forest model using the generated synthetic data to predict patient mortality and observe the results. This shows that synthetic data can solve privacy concerns and still be valuable for medical research.

1. 연구 배경

질병의 진단과 치료는 환자의 과거 의료 기록을 분석하여 체계적인 진단과 치료를 제공할 수 있도록 발전하였다. 과거에는 종이에 수기로 작성하여 보관과 공유가 어렵고 통합적으로 관리되지 못했다. 반면 현대에는 디지털 기술의 발달로 전자 건강 기록(EHR, Electronic Health Records)이 도입되면서 인적 정보, 병력, 약물 복용 정보 및 알레르기, 검사 정보(혈액 검사), 영상의학 이미지, 청구정보 등을 체계적으로 관리할 수 있게 되었다. 이는 의료진이 체계적이고 정확한 의사결정을 내리고, 이를 통합적으로 관리하는 것을 가능하게 하였고, 연구자가 데이터에 더 쉽게 접근하고, 인공지능 기술을 통해 환자의 데이터를 분석하는 등 의료 혁신이 가속화하는 주요한 역할을 한다.

대표적인 EHR 데이터로 MIMIC (Medical Information Mart for Intensive Care)이 있다. MIMIC은 2001년부터 2012년까지 Beth Israel Deaconess Medical Center의 critical care units에 입원한 약 60,000명 환자의 비식별화된 데이터베이스다 [1]. 이 데이터베이스는 Demographics, Vital sign measurements, Laboratory test results, Procedures, Medications, Caregiver notes, Imaging reports, and Mortality 정보를 포함하고 있다. 이를 통해 전세계의 많은 연구자들이 질병 예측, 치료 효과 분석, 재입원 예측, 사망 예측 연구 등에 활용하고 있다.

하지만, EHR은 실제 환자의 정보를 저장하고 있는 만큼 환자의 프라이버시 보호, 법적 규제 준수, 데이터 악용 방지, 그리고 개인정보 보호의 책임이 따른다 [2]. MIMIC 등 대부분의 EHR 데이터는 HIPAA(Health Insurance Probability and Accountability Act) 미국 개인 건강 정보보호 법률에 따라 환자를 특정할 수 없도록 비식별화가 되어 있다. 하지만 비식별화된 데이터라도 여전히 개인정보 유출의 위험이 존재한다. 예를 들어, Re-identification Attack은 비식별화된 데이터를 외부 데이터와의 상관관계를 통해 개인 정보를 복원할 수 있다 [3, 4].

이러한 문제를 해결하기 위해 연구자들은 원본 데이터가 아니면서 통계적 유사성을 나타내고 데이터 분석이나 모델 학습에 사용할 수 있는 합성 데이터를 만들기 시작했다. 합성 데이터는 실제 개인의 데이터를 포함하지 않기 때문에, 개인정보 문제로부터 비교적 자유롭다 [5]. 최근 이미지 생성에 강력한 성능을 보여준 생성형 인공지능 기술인

DDPM (Denoising Diffusion Probabilistic Model)이 테이블 데이터를 생성하는데도 효과적이라고 알려져 있다 [5, 6, 7]. DDPM을 활용해 생성된 합성 데이터는 실제 데이터와 유사한 통계적 특성을 가지며, 다른 EHR 생성 모델의 합성 데이터에 비해 환자의 사망 예측에서 더 좋은 성능을 보였다. 합성 데이터는 개인정보 문제를 해결하면서도 다양한 연구와 분석에 활용할 수 있다 [5].

본 논문에서는 EHR 합성 데이터 생성을 위해 시간-독립 (Time-invariant) 변수와 시간-종속 (Time-dependent) 변수를 모두 처리할 수 있는 모델을 사용한다. 각 타입의 데이터를 각각 임베딩 벡터로 변환 후 결합하고, 이를 DDPM 학습에 사용하여 노이즈로부터 가상의 EHR 데이터를 생성한다.

2. 연구 방법

데이터 수집 및 전처리: PhysioNet에서 제공하는 MIMIC-III 데이터 셋과 eICU 데이터 셋에 FIDDLE 전처리 프레임워크를 적용하여 사용한다 [1, 8]. FIDDLE은 pre-filter, transform, post-filter라는 총 세 단계를 거친다. Pre-filter 단계에서 설정한 범위(T)를 벗어난 데이터와, 일정 비율 이상의 결측치를 가진 데이터를 삭제한다. Transform 단계에서는 시간-종속 변수를 frequent와 non-frequent 변수로 나누고, frequent 변수에 대해서 carry-forward-imputation 적용과 최소값, 최대값, 평균값 통계를 정보를 추가한다. 그리고 모든 카테고리 변수에 대해 one-hot 인코딩을 적용한다. 마지막으로 post-filter 단계에서 대부분이 같은 값으로 존재하는 비정보성 변수를 제거하고 상관관계가 높은 두 변수를 하나의 변수로 통합한다. 본 논문에서는 MIMIC-III에서 변수를 선별, 18세 미만 환자 제외, T=48(hours)로 설정한 논문의 설정을 참고하였다 [9].

생성형 모델 아키텍처: EHR 데이터는 시간-독립 변수와 시간-종속 변수를 동시에 가지는 경우가 대부분이다. 이를 한 모델에서 생성하기 위해 두 개의 독립적인 VAE (Variational AutoEncoder)를 각각의 모달리티에 별도로 사전학습 한다 [10]. 그 다음 DDPM을 학습시키기 위해 각 모달리티를 각각의 VAE 인코더를 통과한 임베딩 벡터를 하나의 임베딩 벡터로 결합하여 입력으로 사용한다. 학습된 DDPM에 노이즈를 입력으로 추가하여 환자에 대한 임베딩

벡터를 생성하고, 각 모달리티에 맞는 디코더를 사용하여 합성 EHR을 생성해낸다 [5]. (그림.1)

표 1. EHR 합성 데이터 생성에 사용된 MIMIC-III 데이터 [9]

MIMIC	
구분	변수
환자 정보	Age, Sex, Pain location, Pregnancy due
임상 검사	Heart rate, Daily weight, Lactate, WBC, NaCl 0.45%, Whole blood, OR urine, Stool, CT scan, X-ray, Sputum, Last dialysis
내원 정보	Admission type, Admission location, Care unit, Ward ID, Admission-to-ICU time

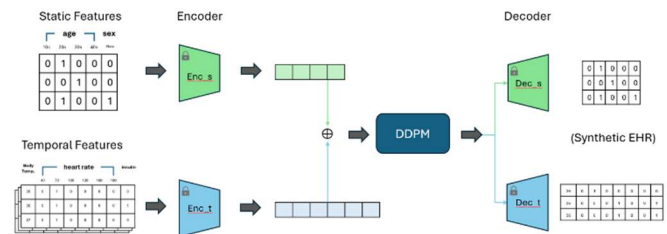


그림 1. FlexGen 모델 개념도

모델 세팅: MIMIC-III와 eICU 데이터 셋을 분리하여 모델을 학습한다. 시간-종속 변수의 경우 시간에 따른 평균 값을 사용했다. 최적화 함수로 Adam을 사용하고, 손실함수로 MSE (Mean Square Error)를 사용하여 학습한다. 학습 완료 후에 사망자와 생존자 합성 데이터를 1,000명씩 생성한다.

생성모델 정확도 평가: 생성 데이터의 정확도는 가우시안 커널 기반의 MDD (Maximum Mean Discrepancy)을 이용하여 원본 데이터와 합성 데이터의 분포 차이를 비교한다.

생성모델 임상적 효용성 평가: 모델의 유용성을 평가는 중환자실에 입원한 환자가 48시간 내에 사망 여부를 예측하는 랜덤 포레스트 모델을 학습하고 정확도와 AUC를 통해 성능을 평가한다.

3. 연구 결과

합성 데이터로 학습한 환자 사망 예측 모델의 성능은 표.2에 정리했다. VAE와 FlexGen 모델을 사용하여 비교 한다.

생성 데이터의 정확성과 임상적 효용성 평가: VAE와 FlexGen를 통해 생성된 데이터는 원본 데이터와 평균적으로 MDD 1.21 차이를 보였다. 특히 FlexGen 모델의 경우 VAE에 비해 0.025 더 높은 정확도와 0.06 더 높은 AUC 그리고 0.07 더 낮은 MDD를 보였으며, 이는 생성된 합성 데이터의 품질이 높다는 것을 의미한다.

4. 결론

본 연구에서는 EHR 생성 모델의 정확도와 임상적 효용성을 평가했다. 시간-독립 변수와 시간-종속 변수를 모두 처리 할 수 있는 모델을 사용하여 생성된 합성 데이터를 통해 사망 예측 모델의 성능을 비교 분석했다. 연구결과, VAE 하나만 사용했을 때 보다 DDPM을 함께 사용했을 때 더 나은 성능향상을 보였다.

하지만 제안된 방법은 시계열 데이터를 시간에 걸친 전체 평균을 사용하기 때문에 시계열 EHR 데이터를 완벽히 재현하지 못하는 한계가 있다. 이는 이 모델을 통해 생성된

합성 데이터는 종단 연구에 적합하지 않음을 의미한다. 따라서 합성 데이터를 실제 임상 연구에 활용하기 위해서 환자의 다중 방문과 시간 축에 따른 변화까지 생성할 수 있는 시계열 모델을 연구할 필요가 있다.

표 2 합성 데이터를 이용한 사망 예측 정확도와 AUROC

	MIMIC-III			eICU		
	ACC.	AUC	MDD	ACC.	AUC	MDD
*(Real data)	0.88	0.83	-	0.88	0.75	-
VAE	0.60	0.62	1.29	0.68	0.72	1.21
FlexGen	0.66	0.72	1.19	0.67	0.74	1.17

*Real data는 실제 데이터를 의미하며, 생성 데이터의 벤치마크로 사용.

5. Acknowledgements

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구 재단(NRF-2022R1F1A1068529), 정보통신기획평가원 (No.2019-0-01842, AI 대학원 (GIST), No.2021-0-02068 (AI 혁신허브)의 지원을 받아 수행된 연구임. 또한, GIST SCENT를 통한 HPC-AI 개방형 인프라의 GPU 컴퓨팅 지원에 감사드립니다.

6. 참고 문헌

- [1] A. P. T. S. L. e. a. Johnson, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, p. 160035, 2016.
- [2] A. Shenoy and J. M. Appel, "Safeguarding Confidentiality in Electronic Health Records.," *Cambridge Quarterly of Healthcare Ethics*, vol. 26, no. 2, pp. 337-341, 2017.
- [3] Yoon, J., Mizrahi, M., Ghalaty, N.F. et al., "EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records.," *npj Digit. Med.*, vol. 6, p. 141, 2023.
- [4] Narayanan, Arvind; Shmatikov, Vitaly, "Robust De-anonymization of Large Sparse Datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, Oakland, CA, USA, 2008.
- [5] Huan He, William hao, Yuanzhe Xi, Yong Chen, Bradley Malin, Joyce Ho, "A Flexible Generative Model for Heterogeneous Tabular EHR with Missing Modality," *International Conference on Learning Representations*, 2024.
- [6] Jonathan Ho, Ajay Jain, Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, p. 6840-6851, 2020.
- [7] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, Artem Babenko, "TabDDPM: Modelling Tabular Data with Diffusion Models," *Proceedings of Machine Learning Research*, pp. 17564-17579, 2023.
- [8] Pollard, T., Johnson, A., Raffa, J. et al., "The eICU Collaborative Research Database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, 2018.
- [9] Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, Jenna Wiens, "Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1921-1934, 2020.
- [10] D. P. Kingma and M. . Welling, "An Introduction to Variational Autoencoders," *arXiv: Learning*, vol. 12, no. 4, pp. 307-392, 2019.