Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230136

Domain Knowledge-Driven Generation of Synthetic Healthcare Data

Atiye Sadat HASHEMI [a], Amira SOLIMAN [a], Jens LUNDSTRÖM [a] and Kobra ETMINANI [a,1]
[a] *Center for Applied Intelligent Systems Research in Health, Halmstad University, Sweden*

Abstract. Healthcare longitudinal data collected around patients' life cycles, today offer a multitude of opportunities for healthcare transformation utilizing artificial intelligence algorithms. However, access to "real" healthcare data is a big challenge due to ethical and legal reasons. There is also a need to deal with challenges around electronic health records (EHRs) including biased, heterogeneity, imbalanced data, and small sample sizes. In this study, we introduce a domain knowledge-driven framework for generating synthetic EHRs, as an alternative to methods only using EHR data or expert knowledge. By leveraging external medical knowledge sources in the training algorithm, the suggested framework is designed to maintain data utility, fidelity, and clinical validity while preserving patient privacy.

Keywords. Domain Knowledge, EHR, Synthetic Data, Representation Learning

1. Introduction and Methods

Generating synthetic data, that is not collected from real-world occurrences but is artificially generated, is considered nowadays as an alternative to making data sharable whilst maintaining the constraints of data privacy and sensitivity [1]. In the healthcare domain, in particular, the availability of high-quality synthetic data can open up opportunities to improve healthcare quality and efficiency, policy evaluation, and large-scale biomedical research investigations. State-of-the-art techniques provide methodologies to generate synthetic electronic health records (EHRs) [2]. Yet, several challenges remain due to the longitudinal and temporal aspects, data heterogeneity, sparsity, skewed distributions, ensuring privacy as well as considering clinical knowledge in the process of modeling healthcare data. The focus of this study is to investigate the integration of domain knowledge to support greater patient-centered outcomes that are close to real clinical data by preserving relationships, distributions, predictive capabilities, and patients' privacy.

Figure 1 illustrates the proposed framework which combines the suggested building blocks including representation learning, generative adversarial network (GAN), post-hoc clinical evaluations, and external domain knowledge sources. Medical ontologies (e.g., International Classification of Diseases (ICD), Anatomical Therapeutic Chemical Classification System (ATC), clinical guidelines, Unified Medical Language System (UMLS), and SNOMED CT) provide powerful sources of information for understanding

[1] Corresponding Author: Kobra Etminani, E-mail: Kobra.Etminani@hh.se.

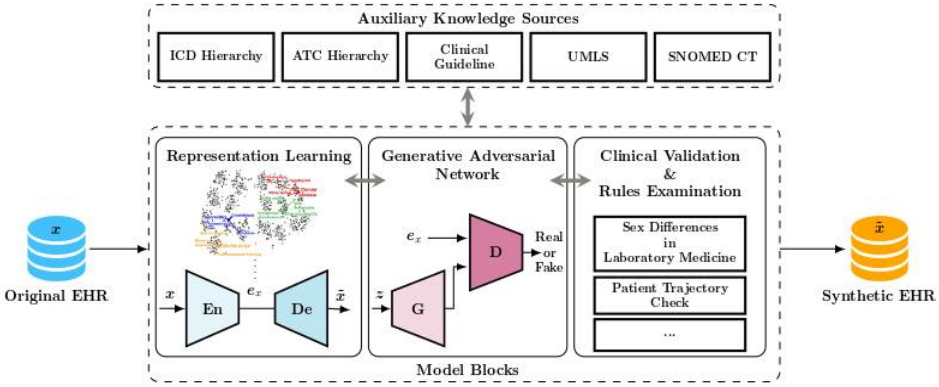

Figure 1. Our proposed data synthesizer framework.

disease progression and enriched categorization of diseases as well as medical treatments. These knowledge sources are integrated for enhancing the conditional representation space for the synthesizer, also allowing us to combine statistical models with data-driven models. The framework uses an Autoencoder as the representation learning module, then a GAN is used for generating data while we have the clinical validity as a ruling-out mechanism. We aim to integrate state-of-the-art representation learning techniques such as graph neural networks (GNNs) and graph convolutional transformers (GCTs) to extract latent structures and relations from EHR data [3].

2. Results, Discussion and Conclusions

While generating realistic data is challenging for researchers, domain knowledge-driven training schemes would be a promising solution. In this paper, we propose a synthesizer that employs a domain knowledge source combined with a deep learning-based model for generating longitudinal EHRs. Although there is some EHR generative software, we aim at considering the clinical knowledge in the synthesizer training process to address the gaps in the multidisciplinary aspects of medical and data science. We hope to outperform previous methods in terms of fidelity and privacy aspects. For this aim, not only do we focus on patients' training data, but also, we leverage diseases-based general knowledge to define several conditions in the training phase (based on the data of respective patient cohorts e.g., heart failure, chronic kidney disease, etc.,) to generate realistic privacy preserved EHRs. In the end, the synthesized EHR will be evaluated via relevant criteria such as fidelity, utility, privacy, and clinical validation metrics.

References

[1] Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: A systematic review. Neurocomputing. 2022 Apr 13.

[2] Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC medical research methodology. 2020 Dec;20(1):1-40.

[3] Chen J, Guo C, Lu M, Ding S. Unifying Diagnosis Identification and Prediction Method Embedding the Disease Ontology Structure From Electronic Medical Records. Frontiers in Public Health. 2022 Jan 20;9:793801.