Chapter 6
Generative AI Approaches for Synthetic Health Data Generation: Methods, Applications, and Challenges

V. Vanitha
https://orcid.org/0000-0003-1010-3241
Sri Ramachandra Faculty of Engineering and Technology, India

M. Srivani
Sri Ramachandra Faculty of Engineering and Technology, India

ABSTRACT

Generating realistic synthetic health data using Artificial Intelligence models offers a powerful solution to privacy restrictions and data scarcity in healthcare. Modern generative frameworks such as Generative Adversarial Networks, Variational Autoencoders, diffusion models, and transformer-based language models enable the synthesis of diverse clinical data, including structured EHR tables, time-series vitals, unstructured clinical notes, and medical images. These models are evaluated on fidelity, utility, and privacy, with applications ranging from data augmentation to privacy-preserving data sharing. GANs excel at learning joint distributions, VAEs capture smooth latent structures, diffusion models yield high-fidelity samples, and transformers produce coherent medical text. Challenges remain in ensuring clinical validity, diversity, and robust privacy protection while avoiding biases and mode collapse. Emerging research explores multimodal generation, promising richer, more integrated synthetic patient records that can transform healthcare research and AI model development.

DOI: 10.4018/979-8-3373-5641-9.ch006

1. INTRODUCTION

Synthetic health data have surged in importance due to privacy regulations and the high cost of collecting large medical datasets. Modern healthcare relies on large-scale electronic health records (EHRs), imaging archives, and genomic databases, but sharing these data is often restricted by laws (e.g., HIPAA, GDPR) or institutional policies. At the same time, many medical applications—such as developing AI models for disease prediction—require abundant, diverse, and richly annotated data.

By mimicking the statistical properties of real data, synthetic datasets enable broader data sharing, data augmentation, and fairness in analysis. Ibrahim et al. (2025) note that synthetic data facilitate sharing while protecting privacy, augment existing datasets, and promote fairness in AI applications. Similarly, Goncalves et al. (2020) emphasize that the scarcity of patient data has slowed progress in medicine, making high-quality synthetic data critical when real data are difficult, expensive, or limited to acquire.

Synthetic datasets can help address the lack of representative samples (e.g., for rare diseases) and accelerate the development of diagnostic models and statistical studies. By mirroring the statistical structures of real data while severing links to individuals, they allow researchers to access realistic datasets without exposing actual patient information. In practice, carefully generated synthetic data support broad sharing, foster algorithm development, and serve as benchmarks—all with minimal privacy risk.

1.1 The Importance of Synthetic Health Data

The key benefits of synthetic health data include:

1. **Data augmentation** – Expanding scarce datasets, especially for rare conditions or under-represented populations. For instance, synthetic images of rare tumors or EHR records of uncommon comorbidities can improve model training and reduce overfitting. Techniques such as Variational Autoencoders and GANs have successfully generated medical images (e.g., pneumonia in chest X-rays) and clinical signals, improving performance on minority classes.
2. **Privacy preservation** – High-quality synthetic EHRs can be shared between institutions without leaking Protected Health Information (PHI). The records mimic real data but are decoupled from real individuals, minimizing re-identification risk, (Nikolentzos et al., 2023).
3. **Fairness promotion** – By oversampling or simulating data for vulnerable or under-represented groups, synthetic datasets can help mitigate bias and improve the fairness of AI models.