Chapter 12
Leveraging Generative AI for Privacy-Preserving Synthetic Data Generation in Healthcare

Ankit Kumar Singh
http://orcid.org/0009-0001-4772-6306
Sri Sathya Sai Institute of Higher Learning, India

Srinath M. S.
http://orcid.org/0000-0003-3275-6999
Sri Sathya Sai Institute of Higher Learning, India

ABSTRACT

Machine learning models require large and diverse data for making accurate inferences. However, large quantities of data are not available in many cases, either due to data privacy issues or rarity of the data. Synthetic data is used to address this requirement as it captures the original distribution and correlation of real datasets. Thereby, guaranteeing high utility by overcoming data scarcity, and privacy challenges. In this chapter, we will discuss the application of generative and foundation models along with privacy-enhancing frameworks that are commonly used to produce realistic and privacy complaint synthetic data. Integrating privacy into synthetic data can dramatically enhance institutional and cross-border cooperation. In addition, the chapter also highlights the necessity of strong evaluation metrics to determine the reliability and domain usefulness of the synthetic data, that provides a scaled avenue of ethical, efficient and privacy-compliant innovation in sensitive healthcare data.

DOI: 10.4018/979-8-3373-7426-0.ch012

1. INTRODUCTION

In the modern digital world, the success of modern Artificial Intelligence (AI) relies completely on access to massive and high-quality collections of information. Much like a student needs to have a library of different textbooks available in order to learn, an AI model needs to get to study vast data sets in order to recognize the complex patterns that it needs to make smart decisions. To use AI in the medical domain. Doctors and researchers now rely on patient variable disease data, clinical text notes, DNA mapping, electronic health record and even real time data from smartwatches. These huge streams of information make "predictive analytics" possible, where computers can detect the signs of illness at early stage or help doctors design a treatment plan that is appropriate for a patient's unique problem condition (Abdalla et al., 2025).

However, it is much more difficult to use this data than it sounds. Even though we have more information than ever before, researchers often run into "bottlenecks" and significant obstacles that slow down their work. For instance, data is often trapped in "silos" meaning it is stored in different computer systems that cannot talk to each other. Furthermore, raw data is hardly ever clean, it is often messy, full of errors or lacks important details. This means that scientists must burn a huge amount of time and efforts "wrangling" and fixing the data before they can even start their analysis. Beyond these technical challenges, there is the important issue of privacy. Due to the very sensitive nature of medical records, there are strict laws and ethical rules that prevent the easy sharing of these medical records.

To solve these problems, we need a major shift in the way we deal with sensitive information. This chapter explores and provides synthetic data generation (Singh, 2025) as a modern solution with essential privacy guarantee. In place of using a real person's private records, one can use AI to produce *"synthetic"* data that behaves exactly same as the real but doesn't belong to any specific individual. By leveraging mathematically secure ways to protect patient's confidentiality.

1.1 The Era of Data-Driven Healthcare and Generative Paradigm

The modern landscape of healthcare and clinical research is undergoing an intense transformation that is fundamentally driven by the aggregation and analysis of huge volumes of information. Large and diverse datasets are the driving force behind personalized medical research and prediction of disease. From high-resolution medical imaging to complex, longitudinal electronic health records (EHRs), the algorithms