Ethical Challenges of Using Synthetic Data

Pavitra Chauhan, Lars Ailo Bongo, Edvard Pedersen

Department of Computer Science, UiT The Arctic University of Norway, Norway
pavitra.chauhan@uit.no, lars.ailo.bongo@uit.no, edvard.pedersen@uit.no

Abstract

There is an outburst of digitized medical data with the growing adoption of Electronic Health Record (EHR) systems but have restricted access due to legal compliances. This lack of data accessibility has piqued the interest in generating and using synthetic data. Synthetic data is programmatically generated using the statistical properties of the real dataset. Although synthetic data tackles the issue of legal compliance, there are some ethical concerns associated with it. In this paper, we discuss three ethical concerns of synthetic medical data such as fairness, privacy and unwarranted use. Further, we identify a need to develop a practical framework for statistical evaluation metrics for synthetically generated data.

Introduction

In the last decade, there has been widespread adoption of Electronic Health Record (EHR) systems by hospitals across geographies. These EHR systems are used to store and manage patient medical records. With the digitization of health records, there is an influx in healthcare application development. These healthcare applications aim to improve healthcare by providing personalized treatment plans, prediction and early disease detection, tracking and monitoring of health parameters through the Internet of Things, and awareness about healthy lifestyles. However, digitized medical records are siloed due to restricted access caused by legal compliances such as General Data Protection Regulation in Europe and Health Insurance Portability and Accountability Act in the US. This lack of accessibility has increased the interest of both healthcare researchers and application developers in generating and using synthetic data.

Synthetic data is generated using machine learning (ML), it statistically captures the original dataset and thus, can be privacy-preserving (Tucker et al. 2020). Recent platforms for generating synthetic data include Syntegra (Mendelevitch and Lesh 2021) and Synthea (Walonoski et al. 2018). To support the development of ML-based healthcare apps, we developed the SyntHIR system (Chauhan et al. 2023). SyntHIR generates interoperable synthetic data which is made accessible to the scientific community for the purpose of research and development (Figure 1).

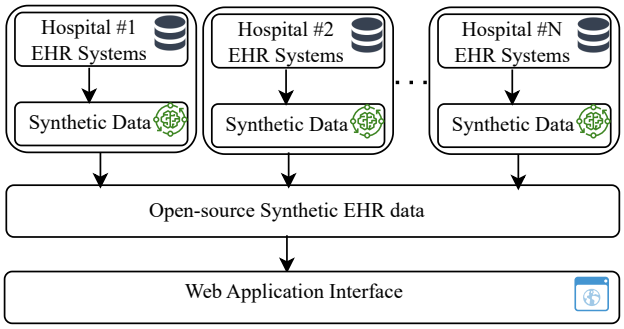

Figure 1: The figure depicts the SyntHIR system design for generating synthetic EHR. It is integrated with different hospital EHR systems and generates synthetic data on-premises in a protected environment. The synthetic data generated from different sources is combined in a single open-source server. This synthetic data is available through a web application interface.

Although synthetic data removes some ethical challenges like informed consent, there are other ethical concerns that we discuss in this paper. However, there is evaluation and fairness associated with the synthetic data generated, which raises ethical concerns for healthcare systems development and healthcare analytics (Arora and Arora 2022).

Ethical Concerns

This section discusses three ethical challenges associated with using programmatically generated healthcare data.

Fairness: Synthetic Data Can Be Biased Unbiased data is an illusion, as decisions such as data selection and analysis are subjective and influenced by perceived importance. Synthetic data is generated based on a real dataset, so patterns and biases in the underlying data will be reproduced. In reality, there is more complexity and nuances that cannot be systematically reflected and accounted for in synthetic datasets. For example, healthcare data is dynamic in nature, and any dataset that forms the basis of generating larger synthetic data will be valid at that point in time. However, even the most statistically relevant synthetic data can grow obsolete as the real data keeps evolving, and the algorithms may

not anticipate the changing factors.

This leads to data-driven bias, which is one of the biggest concerns in the use of Artificial Intelligence algorithms in production (Norori et al. 2021). Algorithms trained on cross-sectional synthetic datasets will become a closed system and will generate a closed and limited set of predictions. Therefore, a mechanism to reduce this "reality gap" is needed.

Privacy: Synthetic Data Can Give an Illusion of Privacy While it is stated that synthetic healthcare data promises privacy, the reality can be different. If the source of the dataset is too small compared to the dimensionality, it may still be possible to infer personal information. It is believed that synthetic data modelled from real data retains enough information to provide protection against various attacks associated with deanonymization and re-association with real patient records. There are no robust methods to determine if the synthetic data generated is truly anonymous. It is shown that synthetic data can be reverse-engineered to the original information. (Stadler, Oprisanu, and Troncoso 2022) evaluated five synthetic data generative models and found that it was possible to infer individual records and reassociate them with the original people, especially in cases where there are statistical outliers. The authors concluded by saying that if a synthetic dataset preserves the characteristics of the original data with high accuracy and thereby retains data utility, it simultaneously enables adversaries to extract sensitive information about individuals.

Unwarranted Use: Synthetic Data Can Be Used Unethically If synthetic data becomes an alternative to real data, it may be misused by healthcare providers and insurance companies. Health insurance companies collaborate with data brokers to collect personal details like daily health vitals, bills and more. This data is unverified and error-prone. So generating synthetic data on it and making medical assumptions could lead to insurers creating improper pricing plans. For instance, raising rates based on false information. This violates three out of the four medical ethics principles stated in (Dunn and Hope 2018) such as Beneficence: the promotion of what is best for the patient; non-maleficence: avoiding harm to the patients and Justice: patients in similar situations have access to the same healthcare.

(Chiruvella, Guddati et al. 2021) discuss how pharmaceutical companies use patient information to target weaker sections of the population that requires more care and consideration in healthcare. In 2019, Avanir Pharmaceuticals was charged with paying physicians' kickbacks to promote prescriptions of its drug named Nuedexta. The primary targets were long-term care facilities with older patients who may have presented with signs of dementia. However, the drug had no proven use in dementia treatment, and its purpose was clouded by the company's misleading information (Department of Justice USA 2019).

Discussion and Proposed Solution

The fairness of synthetic data can be measured using statistical models. Previous work has evaluated the fidelity of synthetic data. In (Bhanot et al. 2021), the authors have developed two metrics to quantify fairness on the synthetically generated publicly available datasets. There are studies to ensure the privacy of synthetic data. In (Bellovin, Dutta, and Reitinger 2019), the author discusses using a technique called differential privacy in combination with synthetic data to handle the issue of a data leak.

We aim to provide quality evaluation metrics that can be reused to assess the relevance of synthetic data generated by any platform. Our objective is to create a statistical evaluation matrix to validate the fidelity, privacy, and diversity of the synthetic data generated. To tackle the outlined ethical challenges, a framework for detecting and alleviating issues will be developed and implemented in SyntHIR.

References

Arora, A.; and Arora, A. 2022. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. *Future Healthcare Journal*, 9(2): 190.

Bellovin, S. M.; Dutta, P. K.; and Reitinger, N. 2019. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22: 1.

Bhanot, K.; Qi, M.; Erickson, J. S.; Guyon, I.; and Bennett, K. P. 2021. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9): 1165.

Chauhan, P.; Askar, M. G. S.; Fjukstad, B.; Bongo, L. A.; and Pedersen, E. 2023. Interoperable synthetic health data with SyntHIR to enable the development of CDSS tools. arXiv:2308.02613.

Chiruvella, V.; Guddati, A. K.; et al. 2021. Ethical issues in patient data ownership. *Interactive Journal of Medical Research*, 10(2): e22269.

Department of Justice USA, U. 2019. Pharmaceutical Company Targeting Elderly Victims Admits to Paying Kickbacks, Resolves Related False Claims Act Violations. https://www.justice.gov/opa/pr/pharmaceutical-company-targeting-elderly-victims-admits-paying-kickbacks-resolves-related. Accessed: 2023-07-06.

Dunn, M.; and Hope, T. 2018. *Medical ethics: a very short introduction*. Oxford University Press.

Mendelevitch, O.; and Lesh, M. D. 2021. Fidelity and privacy of synthetic medical data. *arXiv preprint arXiv:2101.08658*.

Norori, N.; Hu, Q.; Aellen, F. M.; Faraci, F. D.; and Tzovara, A. 2021. Addressing bias in big data and AI for health care: A call for open science. *Patterns*, 2(10).

Stadler, T.; Oprisanu, B.; and Troncoso, C. 2022. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, 1451–1468.

Tucker, A.; Wang, Z.; Rotalinti, Y.; and Myles, P. 2020. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1): 1–13.

Walonoski, J.; Kramer, M.; Nichols, J.; Quina, A.; Moesel, C.; Hall, D.; Duffett, C.; Dube, K.; Gallagher, T.; and McLachlan, S. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3): 230–238.