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

*doi:10.3233/SHTI230063*

# Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review

Shu-Feng TSAO[a,1], Kam SHARMA[a]. Hateem NOOR[a], Alan FORSTER [b] and Helen CHEN[a]
[a] *School of Public Health Sciences, University of Waterloo, Canada*
[b] *Ottawa Hospital, Ottawa, Ontario, Canada*

**Abstract.** With the recent advancement in the field of machine learning, health synthetic data has become a promising technique to address difficulties with time consumption when accessing and using electronic medical records for research and innovations. However, health synthetic data utility and governance have not been extensively studied. A scoping review was conducted to understand the status of evaluations and governance of health synthetic data following the PRISMA guidelines. The results showed that if synthetic health data are generated via proper methods, the risk of privacy leaks has been low and data quality is comparative to real data. However, the generation of health synthetic data has been generated on a case-by-case basis instead of being scaled up. Furthermore, regulations, ethics, and data sharing of health synthetic data have primarily been inexplicit, although common principles for sharing such data do exist.

**Keywords.** Synthetic data, data governance, data sharing, FAIR, CARE

## 1. Introduction

Health data, especially electronic medical records (EMRs), are often stored in disparate systems and formats, rendering integration and standardization difficult [1-2]. Additionally, health data has been strictly regulated by laws, including the Health Insurance Portability and Accountability Act (HIPAA) in the United States (US), the Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada, and General Data Protection Regulations (GDPR) in the European Union (EU) [3]. Researchers and developers often depend on de-identified, aggregated data to test theories, models, algorithms, or prototypes, but it takes a substantial amount of time and resources to retrieve, aggregate, and de-identify relevant data before it can be used [1-2]. One approach to solve this issue is the creation of realistic, high-quality synthetic health datasets that capture as many of the complexities of the original data sets, but do not include any real patient data [1]. In this scoping review, synthetic data is considered different from deidentified, aggregated data. The latter remains to be a type of real data, whereas the former is completely unreal data created from the real data. For example, the Clinical Practice Research Datalink (CPRD) in the United Kingdom (UK) has created synthetic datasets available for research [4]. The Agency for Healthcare Research and

[1] Corresponding Author: Shu-Feng TSAO, E-mail: s7tsao@uwaterloo.ca.

Quality (AHRQ) in the US also has Synthetic Healthcare Database for Research (SyH-DR) available [5]. Synthetic health data can reflect the characteristics of a population of interest and be a useful resource for researchers, health information technology developers, and informaticians. Therefore, health synthetic data provides great promises to protect patient privacy, diversify datasets, and enhance medical and innovative research. Unlike UK and US, Canada has very limited sharable and useful high-quality health synthetic datasets that meet findable, accessible, interoperable, and reusable (FAIR) standards, despite its footprint in the Common Infrastructure for National Cohort in Europe, Canada, and Africa (CINECA) projects [6]. Although there are principles such as FAIR and CARE (Collective benefit, Authority to control, Responsibility, Ethics) [7-8], applications or implementations of these principles on health synthetic data have remained limited.

With the advance of machine learning (ML) and artificial intelligence (AI), generation of synthetic data has been extensively studied [9]. Generative Adversarial Networks (GAN), along with its customizations, have been promising methods for synthetic data generation recently [9]. Although GANs still have its limitations, they can preserve privacy of health synthetic data more than conventional statistical methods [9]. Furthermore, federated learning is another promising technique to protect data privacy and security since it can train AI models without exchanging real or synthetic data across multiple nodes or networks. This can prevent critical data compromises, but it has not been optimized and implemented at a larger scale [10-11]. If GANs and federated learning can be used together to generate FAIR or CARE health synthetic data, it will create a robust and optimal health data network to protect sensitive patient data and accelerate health research and innovations [10-11]. Although scholars have thoroughly investigated methods for synthetic data generation [9], other gaps, including data utility and governance of health synthetic data, have not been studied comprehensively. Therefore, this scoping review aimed to better understand the current knowledge in the identified gaps and future directions for the health synthetic data.

## 2. Methods

The scoping review was completed by following the PRISMA Extension for Scoping Reviews (PRISMA-ScR) [12]. PubMed, Scopus, and Google Scholar were used to search not only peer-reviewed journal articles, but also grey literature related to our research. The primary reviewer screened the titles and abstracts of all possibly relevant articles written in English and published between 2012 to December 2022 to determine whether they should be included in full article reviews and retrieved the full articles. Relevant articles were then read, and each paragraph was coded for specific themes (e.g., DG for data governance) by the primary and secondary reviewers. The entire article was then classified by the majority of paragraph themes. The articles were therefore grouped based on their main theme to summarize the main findings. Any discrepancies were discussed and solved by all the reviewers.

## 3. Results

### *3.1. Evaluations of Data Quality, Privacy, and Utility*

Synthetic Data generation models, utility of the generated health data and privacy concerns of that synthesized health data are co-related. Currently there is no industry standard to produce health synthetic data, however, one of the most popular models are the GANs. These models produce robust synthetic data when the real-world data is also robust by identifying trends in the real-world data without overfitting the synthesized data [13]. Overfitting can take place when the data generated is too similar or almost identical to the real-world data. This becomes problematic with privacy preservation as some examples could be synthesized that are too similar to real-world data.

With the recent boom in electronically stored health related data, there has been a proportional increase in concerns about privacy protection [14]. Synthetic health data synthesis is a key factor in elevating stress associated with health data related privacy concerns. Currently, many models exist to synthesize synthetic data; however, GANS are the most popular.

Common use cases of health synthetic data can be assigned into 6 general categories: (1) EMRs [15], (2) health insurance claims [5, 13, 16], (3) Administrative heath data or surveys [13-14, 16-18], (4) bioinformatics [6], (5) medical images [15], and (6) sensor data [19]. Depending on how data processing is done, data in these categories can be treated as longitudinal or cross-sectional in corresponding analyses.

### *3.2. Health Synthetic Governance, Data Sharing, and Ethics*

Compared to deidentified real patient data, health synthetic data have primarily remained as a grey area in corresponding regulations that govern and protect patient privacy, and its generation and sharing have also been done on a case-by-case basis for research. This has raised many legal and ethical questions that have no clear answers yet. Take informed consents for example. Under HIPAA's privacy rule in the US, creating deidentified data is regarded as healthcare operations of a covered entity [20]. Therefore, informed consents from patients are not required even if the deidentified data will function as a database for research [20]. The similar logic applies to EU's GDPR and Canada' PIPEDA [3]. However, health synthetic data is not de-identified data. Instead, they are fake data artificially created if properly generated, but they closely reflect characteristics of real data. Therefore, this brings up a question: should synthetic health data be considered as protected health information (PHI) or human subject, thus needing informed consents and/or research ethics reviews?

Although health synthetic data appear to be promising for health innovations, sharing synthetic data health is not as common as established databases consisting of real data. Some have advocated that health synthetic data should also follow the FAIR principles for data sharing and open science [8, 21]. Additionally, CARE principles have gained attractions when indigenous data are involved [7, 22]. Existing histories regarding the unfair and unethical treatments of indigenous peoples have strained relationships between indigenous peoples and researchers, resulting in policies that limit data sharing [23]. However, this exclusion of indigenous data sets poses a limitation for a field such as synthetic health data as indigenous datasets can inform many of the machine learning models and clinical algorithms used in research and training [23]. The underrepresented sample of indigenous datasets limits the predictive accuracy of machine learning models, which leads to unintended biases and misinformed data decisions for indigenous peoples and their health [23]. Furthermore, data that has been historically available for indigenous peoples tends to focus on negative outcomes. To increase healthcare access and equity for indigenous peoples, a need for accurate indigenous data is necessary [24]. Synthetic health data will help close this gap in knowledge but requires partnerships with indigenous peoples and synthetic data stakeholders to address this limitation. To inform healthcare and data decisions for indigenous peoples, there is strong need for indigenous data sovereignty. One such way to tackle this issue is with data governance that can be implemented in partnership with indigenous peoples [25]. This also further pushes the discussion of indigenous data sovereignty for synthetic health data. It is unclear who would own that data and in which ways indigenous peoples are involved in the data governance process. Nonetheless, the CARE principles can address historical inequities and provide indigenous peoples a platform wherein they have data sovereignty [7, 22].

## 4. Discussion

As shown in this scoping review, the existing literature about synthetic health data governance and evaluations is scarce. This has suggested gaps to be filled in the future. To generate high-quality and useful health synthetic data, it is important to avoid "garbage in, garbage out." Therefore, the quality of original real data is of great importance. Once synthetic health data are created via proper methods, the risk of privacy breaches becomes lower than deidentified, aggregated real data. However, synthetic data hasn't been generated routinely as a means to share data.

Compared to advanced ML techniques to generate high-quality synthetic health data on a case-by-case basis, data governance, including regulations, ethics, and data sharing, for synthetic health data have remained scarce. Researchers have recommended to follow and apply existing legal and ethical governance, as well as common principles for synthetic health data sharing. However, policies still need to be updated accordingly to explicitly indicate whether or not synthetic health data will be governed as human subject data.

## 5. Conclusion

Health synthetic data offers a promising solution to accelerate health research and innovations. However, its generations and uses have not been scaled up. Further research and regulatory guidelines are needed in data governance and quality evaluations.

## References

[1] Kokosi T, De Stavola B, Mitra R, Frayling L, Doherty A, Dove I, et al. An overview on synthetic administrative data for research. Int J Popul Data Sci [Internet]. 2022;7(1). doi:10.23889/ijpds.v7i1.1727

[2] Kokosi T, Harron K. Synthetic data in medical research. BMJ Med [Internet]. 2022;1(1):e000167. doi:10.1136/bmjmed-2022-000167

[3] Shapiro J. Why digital privacy is so complicated [Internet]. Progressive Policy Institute. 2022 [cited 2023 Jan 5]. Available from: https://www.progressivepolicy.org/publication/why-digital-privacy-is-so-complicated/

[4] Synthetic data [Internet]. Cprd.com. [cited 2023 Jan 5]. Available from: https://cprd.com/synthetic-data

[5] Synthetic healthcare database for research (SyH-DR) [Internet]. Ahrq.gov. [cited 2023 Jan 5]. Available from: https://www.ahrq.gov/data/innovations/syh-dr.html

[6] CINECA - common infrastructure for national cohorts in Europe, Canada, and Africa [Internet]. CINECA. [cited 2023 Jan 5]. Available from: https://www.cineca-project.eu/

[7] Carroll SR, Herczog E, Hudson M, Russell K, Stall S. Operationalizing the CARE and FAIR Principles for Indigenous data futures. Sci Data [Internet]. 2021;8(1):108. Available from: doi:10.1038/s41597-021-00892-0

[8] Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data [Internet]. 2016;3:160018. doi:.1038/sdata.2016.18

[9] Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: A systematic review. Neurocomputing [Internet]. 2022;493:28–45. doi:10.1016/j.neucom.2022.04.053

[10] Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: Systematic review and architecture proposal. ACM Trans Intell Syst Technol [Internet]. 2022;13(4):1–23. doi:10.1145/3501813

[11] Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated Electronic Health Records. Int J Med Inform [Internet]. 2018;112:59–67. doi:10.1016/j.ijmedinf.2018.01.007

[12] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med [Internet]. 2018 [cited 2023 Jan 5];169(7):467–73. doi:10.7326/M18-0850

[13] Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The problem of fairness in synthetic healthcare data. Entropy (Basel) [Internet]. 2021;23(9):1165. doi:10.3390/e23091165

[14] El Emam K, Mosquera L, Jonker E, Sood H. Evaluating the utility of synthetic COVID-19 case data. JAMIA Open [Internet]. 2021;4(1):ooab012. doi:10.1093/jamiaopen/ooab012

[15] Reiner Benaim A, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, et al. Analyzing medical research results based on synthetic data and their relation to real data results: Systematic comparison from five observational studies. JMIR Med Inform [Internet]. 2020;8(2):e16492. doi:10.2196/16492

[16] Yale A, Dash S, Bhanot K, Guyon I, Erickson JS, Bennett KP. Synthesizing quality open data assets from private health research studies. In: Business Information Systems Workshops. Cham: Springer International Publishing; 2020. p. 324–35. doi:10.1007/978-3-030-61146-0_26

[17] Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data [Internet]. 2016;3(1):160035. doi:10.1038/sdata.2016.35

[18] Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol [Internet]. 2020;20(1). doi:10.1186/s12874-020-00977-1

[19] Norgaard S, Saeedi R, Sasani K, Gebremedhin AH. Synthetic sensor data generation for health applications: A supervised deep learning approach. Annu Int Conf IEEE Eng Med Biol Soc [Internet]. 2018;2018:1164–7. doi:10.1109/EMBC.2018.8512470

[20] Nass SJ, Levit LA, Gostin LO, Institute of Medicine (US) Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule. HIPAA, the Privacy Rule, and its application to health research. Washington, D.C., DC: National Academies Press; 2009. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9573/

[21] Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Inf Serv Use [Internet]. 2017 [cited 2023 Jan 6];37(1):49–56. doi: 10.3233/ISU-170824

[22] 1Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, et al. The CARE principles for indigenous data governance. Data Sci J [Internet]. 2020;19. doi:10.5334/dsj-2020-043

[23] Boscarino N, Cartwright RA, Fox K, Tsosie KS. Federated learning and Indigenous genomic data sovereignty. Nat Mach Intell [Internet]. 2022;4(11):909–11. doi:10.1038/s42256-022-00551-y

[24] Walker J, Lovett R, Kukutai T, Jones C, Henry D. Indigenous health data and the path to healing. Lancet [Internet]. 2017;390(10107):2022–3. doi:10.1016/s0140-6736(17)32755-1

[25] Love RP, Hardy B-J, Heffernan C, Heyd A, Cardinal-Grant M, Sparling L, et al. Developing data governance agreements with Indigenous communities in Canada: Toward equitable tuberculosis programming, research, and reconciliation. Health Hum Rights. 2022;24(1):21–33. PMID: 35747272; PMCID: PMC9212824.