Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
4(2): 138–163.
DOI: 10.3934/aci.2024009
Received: 10 September 2024
Revised: 30 September 2024
Accepted: 01 October 2024
Published: 28 October 2024

Overview

Finnish perspective on using synthetic health data to protect privacy: the PRIVASA project

Tinja Pitkämäki[1,*], Tapio Pahikkala[1], Ileana Montoya Perez[1], Parisa Movahedi[1], Valtteri Nieminen[1], Tom Southerington[2,3], Juho Vaiste[4], Mojtaba Jafaritadi[5], Muhammad Irfan Khan[5], Elina Kontio[5], Pertti Ranttila[5], Juha Pajula[6], Harri Pölönen[6], Aysen Degerli[6], Johan Plomp[6] and Antti Airola[1]

[1] Department of Computing, University of Turku, Turku, Finland
[2] Faculty of Law, University of Turku, Turku, Finland
[3] Finnish Biobank Cooperative (FINBB), Turku, Finland
[4] Turku School of Economics, University of Turku, Turku, Finland
[5] Faculty of Engineering and Business, Turku University of Applied Sciences, Turku, Finland
[6] VTT Technical Research Centre of Finland Ltd., Tampere, Finland
* **Correspondence:** Email: tinja.e.pitkamaki@utu.fi; Tel: +358449670054.

Academic Editor: Pasi Fränti

Abstract: The use of synthetic data could facilitate data-driven innovation across industries and applications. Synthetic data can be generated using a range of methods, from statistical modeling to machine learning and generative AI, resulting in datasets of different formats and utility. In the health sector, the use of synthetic data is often motivated by privacy concerns. As generative AI is becoming an everyday tool, there is a need for practice-oriented insights into the prospects and limitations of synthetic data, especially in the privacy sensitive domains. We present an interdisciplinary outlook on the topic, focusing on, but not limited to, the Finnish regulatory context. First, we emphasize the need for working definitions to avoid misplaced assumptions. Second, we consider use cases for synthetic data, viewing it as a helpful tool for experimentation, decision-making, and building data literacy. Yet the complementary uses of synthetic datasets should not diminish the continued efforts to collect and share high-quality real-world data. Third, we discuss how privacy-preserving synthetic datasets fall into the existing data protection frameworks. Neither the process of synthetic data generation nor synthetic datasets are automatically exempt from the regulatory obligations concerning personal data. Finally, we explore the future research directions for generating synthetic data and conclude by discussing potential future developments at the societal level.

Keywords: data protection; healthcare; machine learning; differential privacy; synthetic data

1. Introduction

Data is the raw material that fuels evidence-based decision-making as well as research, development, and innovation (RDI). Researchers, product developers, and decision-makers are increasingly focused on individual determinants of health, making aggregate datasets unfit for their needs. However, data usage must not jeopardize people's fundamental rights, such as the right to privacy. The more detailed individual-level information datasets contain, the greater the privacy risks entailed. Therefore, analyzing and sharing individual-level data requires special attention to data protection.

In 2019, Finland enforced national regulation on the secondary use of health and social data (Act on the Secondary Use of Health and Social Data, 552/2019). The political debates and incentives leading to this reform are described by Aula [1]. The new act treated research and development as distinct purposes for personal data processing, setting more rigorous rules for business-oriented development activities. Against this backdrop, we asked whether synthetic data could narrow the gap between aggregate statistics and pseudonymized data to facilitate the availability of individual-level data for company R&D. By mitigating privacy concerns, synthetic health data shows promise in adding flexibility to different stages of product or service development. While the upcoming European Health Data Space (EHDS) Act [2] will introduce changes to the Finnish regulatory framework, establishing private protocols for exchanging data and insights will remain a pivotal target for the health sector on both national and European levels. In addition to companies, also scientific research and education would benefit from having streamlined access to up-to-date, realistic datasets for demonstrations and testing.

Synthetic datasets have already been successfully applied in stroke and cancer research [3], radiology [4], epidemiology [5], and many other medical disciplines [6]. To incorporate the use of synthetic data into RDI processes, we must increase our understanding of the potential use cases and the associated requirements [7, 8]. For example, the quality and privacy criteria for synthetic data may vary greatly depending on whether the data is used for testing system functionalities or formulating preliminary research hypotheses in a trusted environment. Additionally, many alternative methods exist for generating synthetic data [9], and not all of them are designed to protect privacy. Therefore, the privacy implications of synthetic datasets must be evaluated on a case-by-case basis, taking into account the intended context of use [10, 11].

In this paper, we provide practical guidelines for leveraging privacy-preserving synthetic data for health, with an emphasis on data-driven methods for synthetic data generation. The guidelines are applicable to companies, public entities, and research organizations looking to enhance knowledge transfer, co-creation, and data flows within the healthcare ecosystem. This work summarizes our findings from a Finnish academy-industry collaboration PRIVASA (Privacy-preserving AI for Synthetic and Anonymous Health Data, 2021–2024) and features research works from the project.

In the next section, we will introduce the concept of synthetic data. We then proceed to list the different types of use cases for synthetic data with examples from the health domain. The fourth section addresses legal, technical, and ethical questions that emerge when the goal of synthetic data generation is to protect privacy. Finally, we provide a brief outlook on the methods of synthetic data generation and evaluation.

2. The many interpretations of synthetic data

Synthetic data is often described as an artificially generated, statistical representation of *real-world data (RWD)*, yet there is no single, formal definition for synthetic data. The lack of a consistent interpretation is partially explained by the many possible modalities of synthetic data, including text, numeric data, signal data, and images. Another explanation stems from the fact that synthetic datasets have been studied in parallel within the fields of statistics and computer science [12]. In this article, we have adopted the definition provided by Jordon et al. [13]: *"Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)"*. According to this definition, synthetic data is generated for the dedicated purpose of data science tasks. We interpret this in the broad sense, covering all tasks from exploratory mapping to setting up workflows and performing advanced analytics.

Concepts like fully synthetic data and partially synthetic data also carry slightly different meanings in the existing literature. Raghunathan et al. [14] introduced synthetic samples and populations created via multiple imputation, leading to a definition of fully synthetic data as datasets consisting entirely of modeled data points. In this paper, we endorse the view presented by El Emam et al. [15] that even fully synthetic datasets are not automatically free of associations to RWD. In line with Reiter [16] and following Jordon et al. [13], partially synthetic datasets contain both real (original) and modeled attributes. Hybrid synthetic data mixes real and synthetic data points [17]. As partially and hybrid synthetic datasets include real data in an unaltered form, they are less likely to qualify as anonymous (see Section 4). Hence, this paper focuses on fully synthetic data.

While RWD consists of measurements, observations, and experiences from real life, fully synthetic data is based on a mathematical model representing real-world phenomena. The model itself can be anything from a simple regression model to a complex deep learning algorithm, but an informative link to RWD is typically assumed. Setting synthetic data apart from fabricated (as in *imaginary*) or simulated datasets can be more challenging, as these terms are often used interchangeably, even if the underlying data generation processes differ. In general, fabricated datasets (also known as mock data or dummy data) are created with no input from the real world. Simulated datasets can be based on domain knowledge, parameters extracted from RWD, or both. In contrast to synthetic data generation, the parameters in simulation models often describe key points in a process rather than the overall statistical properties of a dataset.

For example, one could generate a list of random numbers to test how information is transferred within a hospital management system. Taking a step toward realism, one could produce simulated events. If the goal was to perform a patient flow analysis for a healthcare organization, summary statistics from the patient registry could be enough to construct a rule-based simulation, even though more advanced simulation models integrate data from multiple sources [18, 19]. For example, Mohiuddin et al. [18] extracted information from electronic patient records and combined it with domain knowledge to analyze care pathways within a sexual health clinic. Finally, the synthetic data approach for enhancing health care delivery could be to identify a relevant subset of real electronic health records (EHRs), use it to develop a mathematical model (e.g., by training a machine learning algorithm), and explore care pathways using artificial yet statistically accurate data to represent individuals receiving treatment [20]. Healthcare providers, companies, and research organizations may have various reasons for using synthetic data instead of or in combination with the real one, data

protection being one of the most common reasons stated (see Section 4).

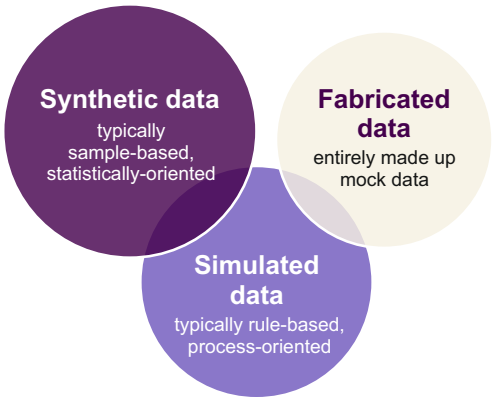

Figure 1. Common properties of synthetic, simulated, and fabricated datasets. The differences are not clear-cut and the terms may be used interchangeably.

In summary, we use the term fabricated dataset in reference to data that is created independently of RWD, whereas simulated datasets are often based on simplified real-world processes. Synthetic datasets, in turn, are generally sample-based model outputs, carefully designed to match the statistical properties of their real-world counterparts (training data) (Figure 1). The use cases for synthetic data tend to put more emphasis on individual data subjects and their attributes in contrast to population-level trends. In the absence of a universal definition, any remarks on synthetic data should indicate the correct interpretation in that context. This can be achieved by specifying attributes like privacy-preserving, fully synthetic, or data-driven, depending on the applicability. Communicating the meanings and underlying assumptions is critical for productive dialogues and fostering trust among different stakeholders, data subjects in particular. In the next section, we will provide an overview of the potential use cases for different types of synthetic datasets before moving on to a more detailed examination of the privacy implications.

3. Use cases for synthetic data: from early tests to private data release

Synthetic data is a promising tool for supporting data analytics workflows, knowledge transfer, and collaboration across industries. The diverse range of use cases (see Table 1) suggests that there is no one-size-fits-all solution, but synthetic datasets should be created to match their intended purposes. Typical reasons for working with synthetic data are

(1) Real-world data does not exist. In this case, generating fabricated, simulated, or synthetic data to test different data analytics workflows could help design a novel data collection protocol. Simulations or synthetic data generation could be based on other existing datasets with properties similar to the phenomenon of interest.
(2) Real-world data exists, but it is not available for intended use. Specifically sensitive health data is subject to strict data protection measures, including access controls and limited purposes of use.
(3) Real-world data exists and is available, but it is of poor quality. Many existing datasets exhibit strong biases or narrow sampling and could be complemented with synthetic data to create a more

balanced dataset [21]. This approach is sometimes referred to as data augmentation.

(4) Real-world data is available and of excellent quality, but making use of it would require investing a lot of resources. Gaining access to real-world datasets can be expensive, or the process may take a long time. It might be a sensible strategy to experiment with synthetic data before deciding on these investments.

(5) Real-world data is available and of excellent quality, but using synthetic data is preferred for ethical reasons. For many applications such as medical care or clinical trials, the most ethical approach is using RWD to produce as accurate results as possible. Other cases are less self-evident. One such gray area could be linked to the data minimization principle in the European General Data Protection Regulation (GDPR, 2016/679 EU), which states that personal data should only be processed when necessary (see also Section 4). Even if using personal data for a given task would be legally compliant, organizations could opt for synthetic data to mitigate the risks of privacy disclosure.

Table 1. Example tasks in which synthetic datasets are already becoming established tools.

Examples of use cases for synthetic data.
1. **Testing and validation tasks:** Synthetic datasets support testing and validating analytical workflows, model performance, or software. Even if RWD was a must-have for the final testing and validation steps, synthetic data could serve as a simple tool for the preliminary inspection of errors. In addition, developers could rely on synthetic data to assess system performance and robustness under novel or unusual conditions.
2. **Machine learning (ML) development:** With synthetic data, ML models can be trained to become more robust and generalizable. One approach is to use synthetic data to complement RWD (data augmentation). If real data is not available, synthetic data could provide a reasonable starting point for ML development.
3. **Exploratory analysis:** New projects may involve the need to design new experimental protocols or data collection campaigns. Having synthetic data available could help researchers understand the structure and common properties of the data, which makes it easier to plan the following steps such as formulating research hypotheses, seeking ethical approvals, or filing the data permit applications.
4. **Demonstration and hands-on training:** Synthetic data could be a valuable asset in educational settings, allowing learners to visualize and experiment with realistic datasets.
5. **Privacy-preserving data sharing:** In privacy-sensitive domains like healthcare and finance, the use of synthetic data could mitigate privacy concerns when data is shared internally or externally. As synthetic datasets are programmed to bear resemblance to the RWD, privacy is not guaranteed by default.

In all the scenarios above, generating synthetic data to work with requires at least some information from the real world. How distinct this background information (training data) can be to produce a useful model output (synthetic data) depends very much on the application. If the goal was to perform an initial impact assessment for planning novel health interventions, one might be able to

generate synthetic data on the city's elderly population based on data from another city with similar characteristics. However, the data utility will be significantly lower if the two cities are distinct in terms of demographics or other aspects affecting the outcome of the intervention. In medical imaging, an increasingly popular research topic is image translation, in which images of another modality are produced based on an available modality [22, 23], such as positron emission tomography (PET) to computed tomography (CT) [24] or magnetic resonance imaging (MRI) to PET [25] translations (see also Table 2 on generating synthetic medical images).

Table 2. Developing generative deep learning-based methods to create indistinguishable, fine-grained synthetic medical images [26, 27].

PRIVASA results: Generating synthetic medical image data with conditional GANs.
Using conditional generative adversarial networks (cGANs), the researchers from Turku University of Applied Sciences created several iterations of image-to-image translation frameworks for MRI, CT, PET and X-ray data. The 3D GAN model successfully generated synthetic multimodal brain MRI images featuring Glioblastoma cancer tumors [27]. Essentially, the GAN model processes two sets of image data: the brain atlas from one patient and the lesion mask information from another. This task is particularly challenging, as integrating lesion information during the synthesis of brain MRIs requires a sophisticated AI pipeline capable of creating realistic tumors in the images. The team also developed deep GAN models to synthesize chest and elbow X-ray images, aiming to enhance image classification tasks [28]. Finally, in collaboration with researchers from the Stanford University and the Turku PET Center, the team developed 2D, 2.5D and 3D GAN methods to generate synthetic standard-dose cardiac PET data from ultra-low dose images [26]. Their approach successfully demonstrated the potential of synthetic data generation in molecular imaging, significantly reducing radiation exposure, scan time and the dose required for patients. However, using GAN-based methods of synthetic data generation also presents considerable challenges, including the risk of hallucinations, difficulties in maintaining high image quality and fidelity, and ensuring that the generated images are both medically applicable and realistic while accurately capturing complex anatomical structures.

For each use case, the potential data analytics solutions can be based on RWD only, synthetic data only, or several ways to combine them. Movahedi et al. [29] showed how machine learning models can be trained on synthetic data and privately tested on RWD (see Figure 2). Given that synthetic data generation is essentially a process of modeling, one must consider which properties of the training data need to be preserved in the synthetic data. In a case study conducted by VTT Technical Research Centre of Finland and the Wellbeing Services County of Southwest Finland, synthetic datasets performed reasonably well compared to RWD in predicting ischemic stroke occurrence (see Table 3). This aligns with the results obtained by Benaim et al. [30], who reported that using synthetic data led to similar conclusions as using RWD in five clinical studies. For certain tasks, such as software testing, conveying structural similarity could be more important than matching statistical properties. Especially in these types of use cases, synthetic data comes close to simulated and fabricated data, and deciding on the best approach relies on case-specific consideration. The potential criteria for identifying the right type of test data could be simplicity, repeatability, and whether there's a need for privacy guarantees.

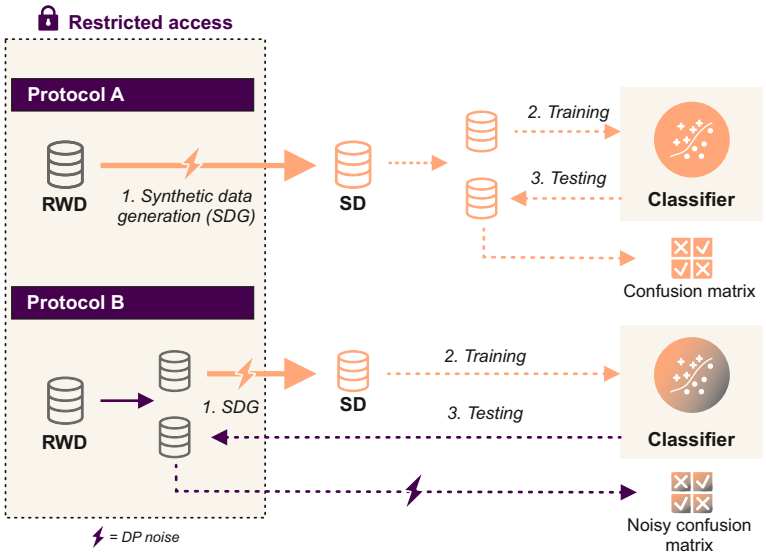

Figure 2. PRIVASA study on how ML models inferred from DP synthetic data generalize to RWD, and how to assess this without compromising privacy [29, 31]. Protocol A uses synthetic data only. Protocol B uses synthetic data for training but conducts testing and validation privately using RWD on the curator's side. Redrawn from Movahedi et al. [29].

Table 3. Evaluation of using synthetic data in place of real patient data to train ML models to predict the occurrence of ischemic stroke.

PRIVASA results: Methods for generating synthetic data for predicting ischemic stroke occurrence.
Using a dataset from the Wellbeing Services County of Southwest Finland, researchers from VTT tested several methods for generating synthetic tabular data on patients diagnosed with ischemic stroke (IS). Both statistical and neural network approaches were represented in the selected methods from the Synthpop data synthesis library [32] and Synthetic Data Vault library (CTGAN, FASTML, Gaussian Copula, and CopulaGAN) [33]. Each method was used to generate 20,000 samples to train the predictive models (Extreme Gradient Boosting, Decision Tree, and Support Vector Machine). This information available to the predictive models represented information from the time before the IS diagnosis. Model performances were calculated using real-world test data excluded from synthetic data generation. The accuracy of models trained with synthetic data was compared to those trained with real data. Higher area under curve (AUC) scores were generally achieved with real training data. However, in many cases, comparable performance was achieved with synthetic training data. For example, XGBoost trained with real data achieved an AUC score of 0.76, and the respective score for Synthpop-generated synthetic data was 0.74. Apart from Synthpop, all methods produced less than 10% of synthetic records with too close of a match with the real data. Our results support the value proposition of synthetic data in initial model training, although real data would still be needed in further stages of model development. This use case also highlighted the importance of having synthetic datasets with measurable levels of privacy protection and quality.

In conclusion, synthetic data serves various purposes across industries and applications. We wish to highlight that synthetic data does not remove the need for real data in developing safe and effective medical treatments, health interventions, or technological innovations. We see privacy-preserving synthetic data as a tool that allows organizations to experiment with new ideas, support decision-making, and build data literacy. In most cases, synthetic data represents an intermediate step to enable the first stages of research or development work and helps identify potential challenges early on. All-purpose synthetic datasets mainly hold value for demonstrative purposes, such as teaching or creating prototypes. For training machine learning models and statistical inference, synthetic datasets should be carefully crafted to reflect the RWD attributes that are considered most relevant for the use case. Finally, synthetic data should not be relied on when achieving real-world accuracy is of the highest priority, such as in clinical decision-making.

4. Synthetic data for preserving privacy

Data protection covers a set of principles for safeguarding personal data against unauthorized access and use. Regulatory frameworks such as the GDPR (2016/679 EU) (Table 4) and Health Insurance Portability and Accountability Act (HIPAA, 1996) in the United States determine when collecting, processing, and sharing personal data is justified. Special attention must be given to sensitive data, including information on personal health. In this context, it is critical to highlight that synthetic data is not anonymous by default. When applied accurately, generative models can indeed produce synthetic datasets with strong privacy-preserving properties, but not without affecting data quality. The well-known privacy-utility trade-off [34–36] introduces a number of technical, legal, and ethical considerations regarding the use of synthetic data as a privacy-preserving mechanism.

4.1. Privacy breaches can take many forms

Any data protection measure should be evaluated against different types of privacy breaches that could occur. The main types of risks recognized in the field of statistical disclosure limitation are

(1) Identity disclosure [15, 20]: An unauthorized person learns that a certain piece of information in the dataset relates to one or more identifiable persons. Synthetic datasets provide effective protection against identity disclosures, as long as information from the training data is not matched too closely (overfitting). Specifically, unique outliers increase the risks, as recognizable data points could make it possible to trace back the properties of real training data and gain additional information.
(2) Attribute disclosure [20, 37]: An unauthorized person uses the data to discover new information about someone they already knew something about. Personal data does not need to be based on real, accurate, or true records. For example, synthetic data could provide a strong indication that someone has a rare medical condition given their very specific diet which is known to the co-workers. This does not make the synthetic dataset itself personal data, but only the information (attribute) that becomes linked with an identifiable person should be considered as such.
(3) Membership disclosure [20, 38]: An unauthorized person learns that an individual or a group of individuals were included in the dataset. In the case of synthetic data, membership disclosure could be interpreted as being included in the real-world dataset that was used to train the generator. Depending on the context, membership disclosure can be immediately followed by an attribute

disclosure. This happens, for example, if the dataset is known to cover patients with a specific medical condition.

In the past, data anonymization implied eliminating all variables that directly pointed out to an individual, including name, social security number, and home address. Since then, it has been recognized that successful identification is possible even with seemingly general and nonpersonal pieces of information. According to influential research by Sweeney [39], approximately half of the US population could be identified based on three data points alone: their place of residence, gender, and date of birth. Therefore, deleting the obvious (direct) identifiers is not enough, and deleting the less obvious (indirect) ones is practically impossible, as almost any attribute could become an indirect identifier when combined with additional information. This is the problem of anonymity that the traditional anonymization methods struggle to overcome.

Approaches like k-anonymity [40], l-diversity, and t-closeness [41] may prove insufficient, if anonymized data is linked to external sources of information (linkage attack) or subjected to intricate analyses to identify persons based on statistical likelihoods (inference attack). Synthetic data offers stronger privacy protection against these types of attacks, because one should not be able to connect artificial data points and RWD with a high level of confidence. Different methods, such as differential privacy (DP) [42], could be applied to ensure that synthetic data generation does not simply copy-paste training data but produces novel data points – mimicking, but not replicating the RWD.

4.2. Personal data processing to generate synthetic data: Accessing real-world health data under the Finnish regulatory framework

When synthetic data is generated through a process of modeling, it is critical to assess the privacy implications of both input and output data. Creating high-quality synthetic health data typically requires, or has at some point required, personal data to ensure a close enough resemblance between synthetic data and RWD. Even if the model output could be considered nonpersonal (privacy-preserving synthetic data), the same might not be true for the model input. On the contrary, generative models are often trained with pseudonymized datasets, in which case the requirements imposed by the data protection regulations still apply to the training stage.

When synthetic data generators are trained with real patient or social care data, the required input data can be subject to additional professional secrecy rules. In Finland, the relevant legislation comprises mainly of the GDPR (2016/679 EU), the Data Protection Act (1050/2018), and the Secondary Use Act (552/2019). Depending on the dataset, also the Biobank Act (688/2012) may apply. Accessing patient records to create synthetic data requires permission in accordance with the Secondary Use Act. The national data permit authority Findata* will decide on permissions in case of 1) records from private health care providers, 2) records of several public health care providers, or 3) records from health care providers that have transferred their decision-making powers to Findata. In other situations, permissions are available via the public health care provider whose patient data is required.

In terms of getting access to the training data, two potential complications arise out of the Finnish Secondary Use Act: individual-level data is only available for scientific research and processing within an audited data processing environment, and approved secure operating environments like Findata's

*https://findata.fi/

own Kapseli support limited data formats and data processing capabilities, which makes them less suited for tasks like ML development. However, only aggregated statistical data can be processed outside these environments. No individual-level patient or social care data is available for pure product development, regardless of where it would be processed.

Patient data is available also from the Finnish biobanks under the Biobank Act, but access requires that the data is accompanied by a biological sample or sample-originating data, such as genomic data. Data from the six Finnish hospital biobanks and the biobank of the Finnish Institute for Health and Welfare (THL) can be applied with a single application through the Fingenious portal[†], whereas the other biobanks can be contacted individually. Contact details are available from the Finnish Medicines Agency Fimea[‡]. Data from biobanks is available for both scientific research and product development and does not need to be constrained to audited and approved processing environments.

The Finnish Secondary Use Act and the Biobank Act enable access to data for a defined project for a defined time, which may pose a challenge for any projects or systems requiring long-term access to the original data. Neither act requires project-specific consent from data subjects.

Possibilities to access data that are not subject to the Secondary Use Act or Biobank Act should be evaluated against the data protection rules and sector-specific legislation, if there are any. For example, data collected as a part of providing health or activity-related consumer services could be accessed to create synthetic data based on consumer consent or other GDPR and Data Protection Act compatible legal bases, remembering the additional requirements for health and other sensitive data under the GDPR Article 9. The GDPR transparency requirements need to be respected, ensuring that the data subjects are kept informed of personal data processing and their rights. Before any personal data processing begins, the privacy risks should be carefully evaluated and the rights of data subjects acknowledged, in case a patient chooses to withdraw their consent or object to processing their personal data.

4.3. Synthetic data as anonymous data: Meeting the GDPR requirements

Using synthetic data instead of real data satisfies the GDPR data minimization principle (Article 5.1c) and exempts processing from data protection requirements, provided that the synthetic data no longer constitutes personal data. The GDPR concept of personal data is very broad, but also elusive, relative, and context-sensitive [43, 44]. In accordance with the definition, personal data means any information relating to an identified or identifiable natural person. The definition requires further clarification of what is information, what does "relate to" mean, and who are natural persons. Recital 26 provides guidance for interpretations, stating that when considering identifiability, all means should be taken into account that could reasonably be used to identify persons. It is notable that the GDPR includes a reasonability test and does not require absolute universal anonymity for data to not be personal (Table 4). To be considered anonymous, it seems clear that synthetic data should not have data points with one-to-one linkages to the real (original and thus identifiable) dataset. It should not be generated by such recoverable methods that could be used to rediscover original data, exposing data subjects to a risk of identity disclosure (model inversion attack) [45].

[†] http://www.fingenious.fi
[‡] https://fimea.fi/en/supervision/biobanks/national-biobank-register

Table 4. A list of general misconceptions on GDPR.

General misconceptions on GDPR.
1. **GDPR requires consent:** The GDPR requires a legal basis for processing personal data, and consent is just one of the possibilities. In many cases, consent may not be the best option or even a feasible one. For example, the actual or perceived power imbalance could prevent public healthcare authorities from obtaining data subject's consent that would be valid in the eyes of the data protection authorities. To be valid as a legal basis, consent must be "freely given, specific, informed and unambiguous" (Article 4).
2. **GDPR requires anonymity:** The GDPR requires a valid legal basis for processing personal data, additional justification for processing sensitive personal data, and data minimization. Depending on the purpose, the data can be directly identifiable, indirectly identifiable (de-identified or pseudonymized), relatively anonymous, or universally anonymous. From the GDPR perspective, "anonymous" implies that there are no means that would reasonably likely be used to identify a person (to link the data to an identifiable person). Based on the European Court of Justice ruling (Breyer, Case C-582/14), the same data can be personal data to one party and not personal data to another, depending on whether or not they have legal access to additional identifying data.
3. **Synthetic data is anonymous data:** Synthetic data can be personal data; see discussion above.
4. **De-identified is anonymous:** De-identification typically means removing certain variables from the dataset, including names, personal IDs, and addresses. In the United States, the HIPAA lists information that needs to be removed for data to qualify as de-identified. An alternative method is to have an expert evaluate the likelihood of identification. HIPAA allows de-identified data to remain indirectly identifiable through coding and code-keys, i.e., pseudonymization. From the European perspective, de-identified data may or may not qualify as anonymous data.
5. **Pseudonymised data is anonymous data:** In Europe, pseudonymization is a common safeguard to protect personal data and minimize the use of directly identifiable data, but pseudonymized data is still understood to be personal data and the same requirements still apply.

A framework that has gained popularity over recent years is differential privacy (DP) [42], which is based on carefully calibrated noise. Adding this noise during the training stage guarantees that the synthetic data generator produces outputs with minor deviations compared to the RWD (see, [46–48]). These deviations should occur at the level of individual observations and in a way that preserves the main statistical properties of the training data. As a result, inferring any personal information from synthetic data becomes difficult due to the low level of confidence or, in other words, an increased margin of error. Concealing personal data with random noise makes it extremely difficult to bypass the data protection, but the cost is paid in data quality. When the amount of noise added is high enough, even the most robust statistical patterns become obscured, and new, artificial patterns could emerge by chance alone [49].

This leads to two practical questions:

(1) How much noise is enough to protect personal data?
(2) How much noise will destroy the data utility?

The answers depend on the task at hand, but, generally speaking, smaller datasets with weaker statistical trends are more sensitive to the DP noise. If the phenomenon of interest is very complex and multidimensional, applying even a small amount of DP noise may break down the multivariate correlations, rendering data useless for hypothesis testing. The data could still serve its purpose in the exploration and building of data analysis pipelines, highlighting the context-specific definition of utility. The need for random noise to conceal personal data could be minimized by using synthetic data in combination with other privacy-enhancing technologies such as federated learning (Table 5).

Table 5. Improved weight aggregation methods for federated learning [50–52].

PRIVASA results: Federated learning as a privacy-preserving alternative to synthetic data.
In federated learning, data is not shared but collaboratively used by training ML models in a decentralized manner [53]. There is no need to transfer the data to a central server because each dataset is processed locally. The results from these local models (e.g., model gradients) are then shared to create a global model. Researchers from the Turku University of Applied Sciences focused on how the information from local models could be combined to produce the most accurate results without leaking any sensitive information from the original, locally processed datasets. They published novel methods, similarity weighted aggregation (SimAgg), regularized aggregation (RegAgg), and regularized SimAgg (RegSimAgg) [50–52], which were subsequently developed further. The team's work was awarded top positions in the International Federated Tumor Segmentation Challenge in 2021 and 2022. All in all, federated learning is a promising approach to joint analytics tasks that require combining real-world data from multiple sources. Federated learning can also be applied for collaboratively training generative models to create more representative synthetic datasets.

4.4. Ethical questions emerging from the use of synthetic health data

As privacy-preserving synthetic data is often algorithmically generated, and possibly used in AI development, the ethical aspects of synthetic health data reflect the wider discourse of AI in healthcare. Privacy-preserving synthetic data could facilitate ethically aligned data sharing practices, one of the most popular research topics being data augmentation to mitigate bias [54]. Yet it also raises complex and profound questions on data ethics, introducing unique viewpoints that are not fully addressed by the existing ethical frameworks [55]. This happens because applying synthetic data in healthcare requires balancing many aims, sometimes conflicting, such as ensuring data confidentiality and making data accessible for the public good.

The ethical questions stem from trade-offs involving data privacy, data quality, fairness, and transparency (Table 6). One could ask questions such as: When is it acceptable to compromise on data quality to protect privacy? Is it ethically appropriate to increase the representativeness of a given dataset with synthetic data? Could synthetic data amplify bias? Should synthetic data generation

always include bias and stability assessments, even if those consume the privacy budget and therefore increase the risk of privacy disclosures? As with legal interpretations, there are no simple answers. Several authors have already pointed out that synthetic data is not free of risks, but rather a tool that reshapes how different types of risks are manifested [56, 57]. For example, Ganev et al. [58] showed how the downstream analyses of DP synthetic datasets provided less accurate or consistent results for minority groups. At the same time, synthetic datasets can promote explainability by helping researchers map model outputs with different types of input data [59].

Compared to other forms of data-driven research, synthetic data is more frequently relied upon as the best out of limited options. For example, RWD could be completely unavailable or of extremely poor quality. Therefore, the ethical assessment should not be based on comparisons to RWD alone, but also acknowledge the "no data" scenarios when applicable. The ethical assessments tailored for synthetic data generation and utilization remain a developing area. However, several published guidelines on responsible and trustworthy AI already provide a high-level framework for identifying best practices.

Table 6. Ethical challenges associated with the generation and use of synthetic health data [60].

PRIVASA results: Ethical implications of synthetic health data.

1. **Human agency and oversight:** Even if synthetic data preserves privacy, the indirect involvement of human participants in data generation calls for clear communication and adherence to individual rights. As discussed by Whitney et al. [57], informed consent should not be disregarded in the context of synthetic data generation.
2. **Technical robustness and safety:** Any methods for synthetic data generation should be evaluated against context-sensitive thresholds of accuracy and reliability.
3. **Privacy and data governance:** Synthetic data may still be susceptible to privacy breaches, which highlights the need for holistic data protection impact assessments (DPIAs).
4. **Transparency:** To build trust, synthetic data should be generated and used in a transparent manner. For example, data labels and model cards could support communicating the origin and limitations of synthetic datasets or models trained on synthetic data.
5. **Diversity, non-discrimination, and fairness:** The process of synthetic data generation should involve meticulous data preparation, use of diverse and representative datasets, and in-depth analyses of data generation algorithms to address biases.
6. **Societal and environmental well-being:** The use of synthetic medical data should not compromise any ethical principles in pursuit of abstract notions such as 'public good' or 'technological progress'.
7. **Accountability:** Actors and organizational entities accountable for the organization's existing data practices should also assume liability when it comes to the generation and utilization of synthetic health data.

For example, to prevent bias in synthetic health data, one could work on the following aspects:

(1) Ensuring the diversity and representativeness of RWD training data.
(2) Observing potential biases during model development.

(3) Applying fairness metrics to synthetic data or generative model [61–63].
(4) Testing the robustness with adversarial training techniques designed to reveal any vulnerabilities.
(5) Matching the data properties with the intended use and making sure all users are aware of the limitations.
(6) Involving experts from different fields to understand all the potential sources of bias (ranging from technical ones to individual behaviors and differences in organizational practices).
(7) Prioritizing transparency and explainability throughout the different stages of synthetic data generation.

A conservative ethical argument could be that every use of synthetic data needs a separate ethical analysis and consideration, especially from the representativeness perspective. In addition, it will be important to establish "digital chains of custody" for synthetic datasets to document how the data was formed, communicate limitations, and track usage [54]. Synthetic datasets should be clearly marked as such and unintentional mixing with real data should be avoided. Progress on the synthetic data front could still be considered a positive direction, as it can diversify and complement the existing data protection protocols.

4.5. Recommendations for privacy-preserving synthetic data

To sum up Section 4, we present the following recommendations for the responsible use of synthetic data to protect privacy:

(1) While privacy metrics provide a practical starting point for employing synthetic data for privacy protection, more work is needed to establish a holistic approach. Based on our experience, releasing synthetic datasets for unrestricted, public use is rarely a realistic goal due to cumulative privacy risks. Instead, data synthesis could be successfully applied with other privacy-enhancing technologies, possibly allowing the other safeguards to be more lightweight than without the synthetic data layer. When data accuracy is the highest priority, secure data analytics protocols are likely to yield better results than sharing differentially private synthetic data.
(2) At present, the legal implications of privacy-preserving synthetic data are not fully resolved, and new frameworks like DP are challenging traditional approaches to statistical disclosure control. As the regulatory practice is forming, forerunner organizations would benefit from public guidelines and benchmarks on how to provide strong enough privacy guarantees. Data protection authorities and governmental organizations worldwide have already taken action, and Finland is well-positioned to coordinate strategic work on a national level.
(3) From legal and ethical perspectives, privacy-preserving synthetic data links to the broader discussion of AI in health care. Avoiding overuse and misuse (*sensu* [64]) is equally important as recognizing the opportunities. We argue that synthetic datasets should be used as a complementary tool to mitigate the current issues on data availability, not as an alternative solution that would replace or restrict current efforts focusing on RWD. For example, using synthetic data to cut costs should not justify lower recruitment rates for clinical trials or less representative sampling in medical studies. Real data should remain at the core of health-related research, and high-quality RWD remains essential for validating any results obtained using synthetic data.

5. Generating synthetic data

Synthetic data generation can be based on domain knowledge (a knowledge-driven process), existing data (a data-driven process), or a combination of both. Synthetic data generated through a knowledge-driven process [65] can resemble simulated data, especially when it comes to synthetic longitudinal data. As a rough distinction, synthetic datasets could be interpreted as snapshots in time, whereas simulated datasets tend to be more process-oriented overall (see Section 2). In this paper, we focus on the data-driven approach.

The computational methods for data-driven synthetic data generation encompass a variety of techniques, from statistical modeling to ML and other AI technologies, including generative AI [9]. Many methods have also been published as privacy-preserving versions, for example, by applying DP guarantees [46–48]. All methods have their strengths and weaknesses, and selecting the best possible one for any given task requires understanding the underlying data as well as the generative model [66]. The selection can be guided by metrics that describe the utility and privacy of synthetic datasets [67,68] or model properties like robustness and explainability [69]. However, any quantitative metrics are inherently imperfect, as they cannot capture the full complexity of technical, legal, and ethical considerations involved.

When it comes to narrowing down the options, the first step is to filter based on the type of data. One main category of methods is formed by generative models for unstructured data (images, video, audio) and the other one for structured data (tabular data, time series). In addition, the methods for generating synthetic tabular data may support only certain types of variables (categorical, numeric discrete, or numeric continuous) [70]. Specifically, the methods for generating synthetic time series are still relatively rare [71], as high-dimensional datasets are the most challenging ones to model without losing the characteristics of RWD. An illustrative overview of generative models for specific data types has been published by Jordon et al. [13].

5.1. The general overview of data-driven methods for synthetic data generation (SDG)

In statistical modeling, data is generated based on the statistical properties of the original dataset. In the simplest form, one could generate new data points based on a predetermined probability distribution. More complex approaches include models like Bayesian networks with conditional probabilities [72, 73]. There is no clear boundary between statistical methods and ML, as Bayesian networks have also been applied in ML algorithms.

AI-driven synthetic data generation covers ML approaches such as GANs [74] and variational autoencoders (VAEs) [75]. GANs involve a dynamic process where two neural networks, a generator and a discriminator, compete to produce realistic data samples (see [76] for a comparison of several GAN-based SDG methods and Table 7 for work conducted in PRIVASA). VAEs, on the other hand, use probabilistic encodings to generate new data points [77]. Large language models (LLMs) excel at producing realistic outputs as free-form text, but can also handle other types of data such as tabular data [68].

Table 8 presents an overview of the data-driven workflow for synthetic data generation. We note that evaluating the amount of data preprocessing, if any, also requires technical expertise: Latner et al. [66] mentioned data cleaning as an essential step before synthetic data generation, whereas Dankar and Ibrahim [78] provided evidence that data synthesis from raw data yields good results. For a practical reference, a paper by Yan et al. [79] includes a tutorial to describe the process of GAN-based synthetic

data generation.

Table 7. Empirical evaluation of a GAN-based method for generating synthetic tabular data [80].

PRIVASA results: Generating differentially private synthetic tabular data with GANs.
The goal of this empirical work was to explore the combined effectiveness of DP and various rates of subsampling in the private generation of synthetic tabular data with GANs. The subsampling increases privacy by dividing the privacy budget across mutually exclusive subsets of training data. This way, each individual data subset will have limited influence on the trained model. The higher subsampling rates were associated with more training iterations before reaching higher model accuracies [80]. The increased cost of training was reasonable considering the privacy gains but had an impact on how the work was conducted in practice (e.g., relying on a single dataset). Generating synthetic tabular data of high quality, however, proved difficult. Some of the effects were observable as artificially reinforced or even reversed correlations.

Table 8. General aspects to consider on different stages of synthetic data generation. The list is indicative.

The process of synthetic data generation.
1. Defining the use case: • Specifying what the synthetic data will be used for and by who. • Conducting an ethical and regulatory assessment. • Defining the utility criteria. • Defining the privacy criteria. 2. Setting up the training data for generative models. • Identifying potential data sources and gaining access to suitable datasets. • Determining the subset of variables that need to be synthesized. • Checking for potential biases and outliers of the RWD. • Checking other potential limitations such as missing values. • Evaluating the need for data preprocessing or augmentation. 3. Building the data synthesis pipeline: • Adjusting technical specifications to match the utility and privacy criteria. • Addressing other use case requirements (legal, ethical, or technical). • Documenting the work and choices made (e.g., code annotations, model cards). 4. Evaluating the generative model and quality of synthetic data: • Model evaluation (e.g., model stability, privacy, and fairness) • Qualitative metrics for synthetic data (e.g., expert opinion) • Quantitative metrics for synthetic data (e.g., general utility, case-specific utility) • Level of privacy protection • Documenting the work (e.g., metadata and data catalogues). 5. Using synthetic data: • Sharing relevant results (synthetic data, generative model, analysis results, metadata). • Sharing supporting documentation, when relevant (e.g., instructions for use).

5.2. *Evaluating the quality of synthetic data*

The quality of synthetic data can be assessed indirectly by looking at the properties of the generator or directly by looking at the properties of the output, i.e., the synthetic data. Several authors have presented their own frameworks for evaluating the quality of synthetic data (see, for example, [81] and [67]) and measuring privacy (see, for example, [82] and [83]). Recent work by Vallevik et al. [84] aims to consolidate the existing evaluation metrics for synthetic tabular data.

The quality of synthetic data is typically described using attributes such as utility, fidelity and privacy [13]. Utility is a measure of how well synthetic data performs in the task it was generated for. In hypothesis testing, one could empirically assess the probabilities of landing a false conclusion (Figure 3) [49]. Fidelity is a measure of how much synthetic data resembles the original data. Privacy metrics describe the level of data protection achieved, usually indicating either overall similarity to RWD or protection against a specific privacy attack.

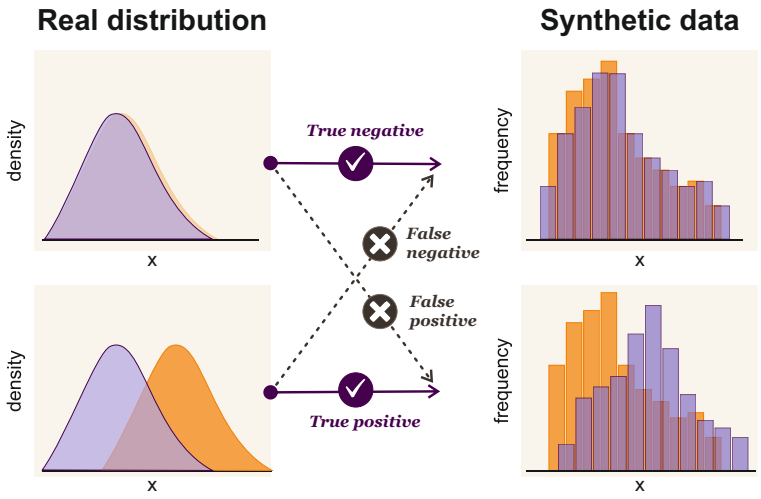

Figure 3. PRIVASA study on assessing the reliability of DP synthetic datasets in hypothesis testing [49]. A simple statistical test that compares two distributions can produce four different outcomes, and false positives (Type I error) or false negatives (Type II error) can lead to misguided conclusions. Redrawn from Montoya-Perez et al. [49].

Utility and fidelity are often tightly interlinked, but sometimes even synthetic data that has little resemblance to RWD can be very useful. Similarly, privacy and fidelity may be hard to separate. As a rule of thumb: the more similar synthetic data is to RWD, the weaker the protection of privacy. Utility is a highly context-specific feature that often requires benchmarking with RWD. The model performance with RWD versus synthetic data can be compared using metrics such as prediction accuracy and overlap (%) of confidence intervals in model coefficients. Fidelity indicates the structural and statistical similarity to RWD by focusing on individual variables, interactions between two variables, or dataset properties at the population level. It can be quantitatively measured by calculating basic descriptive statistics or analyzing distributions. The potential metrics include the structural similarity index (SSIM; for images), pair-wise correlations, and the Hellinger distance or Kolmogorov-Smirnov test for measuring the distance between synthetic data and RWD. Especially with unstructured data like

images and video, a qualitative expert evaluation could be the ultimate test of similarity.

Considering the typical inverse relationship between privacy and utility or fidelity, some of the metrics could be applied in both categories. It must be noted, however, that this inverse relationship is not a simple linear one, but represents a case-specific optimization challenge. The privacy metrics often reflect the distinguishability of synthetic data from RWD (e.g., Propensity score [78]) or the distance between synthetic and real data points. Instead of using these indirect measures, the risk for privacy breaches such as re-identification could also be calculated or empirically tested.

A somewhat contrasting approach is presented by DP [42], which provides a value for the privacy parameter epsilon (ϵ). Lower values translate into stronger privacy protection, which in the case of DP means that the inclusion or exclusion of any individual's data does not significantly affect the observed model output. In contrast to other privacy metrics, DP guarantees apply to the generative model and not the data *per se*. Other properties that could be evaluated at the level of the generative model include bias and stability assessments [30, 67].

5.3. Future research directions for synthetic data generation

We conclude Section 5 with the following recommendations for implementing synthetic data generation techniques and future research:

(1) We support the idea that synthetic datasets should be created for a predefined purpose that serves as the basis for method selection. More empirical research is needed to help evaluate the synthetic data approach against other privacy-enhancing technologies like federated learning [85].
(2) When producing synthetic data, it is important to examine its quality from multiple perspectives. While establishing common standards and benchmarks remains an important goal in the field, employing a diverse set of metrics should be encouraged to avoid setting unrealistically narrow optimization targets. Qualitative evaluations, preferably in collaboration with clinical experts or other healthcare professionals, should not be neglected. The case-specific evaluation frameworks should integrate the relevant quantitative and qualitative aspects, and this information should be made available as metadata.
(3) In line with Hernandez et al. [86], we recognize the need for collaborative, secure protocols for synthetic data generation and validation. Especially in the health sector, we expect an increasing demand for federated synthetic data generation [87].
(4) Data on individual lifestyles has high relevance for preventive healthcare, and it covers many types of data. Future work is needed to generate synthetic data from each of these modalities individually (e.g., trajectory data [88]), but also in ways that support producing multimodal datasets.

6. Conclusions

Synthetic data is often put forward as a solution for data-driven innovation in the face of escalating privacy concerns. As the technology matures, on-demand synthetic data generation could help address different data requirements in research and development, but also knowledge-based management and policy-making in the health sector [89]. Privacy-preserving synthetic data has the potential to support more streamlined access to individual-level health data without compromising privacy. The emerging opportunities for testing and development would benefit company R&D, public-private partnerships as

well as international collaboration. When deciding on the use of synthetic data, one should consider the legal requirements and ethical code of conduct, and assess the technical trade-offs. Where appropriate, these should be compared to the alternative approaches available.

Establishing sufficient levels of data protection remains a challenging task, which forms a barrier to adoption of new privacy-enhancing technologies, including synthetic data. The dichotomy between anonymous and non-anonymous data present in data protection laws is not realistic [90], because data protection forms a continuum with low-risk data at one end and high-risk data at the other. Quantitative metrics such as differential privacy can promote the development of sustainable data strategies and support transparent communication with data subjects. In assessing the overall risk of privacy breaches, however, applying synthetic data generation as a privacy-preserving mechanism represents only one part of the equation. The strength of privacy guarantees could be adjusted based on other safeguards applied, such as restricting the data to be processed within a secure operating environment only.

As with any emerging technology, the wider adoption of synthetic data as a privacy-preserving mechanism also requires research, conceptualization, testing, and reassessment of existing practices. Technological breakthroughs, such as generative AI, should not be seen merely as threats to privacy but also recognized for their potential for enhanced data protection. Enabling legislation has been identified as one of the key areas for development in, for example, recently published growth and competitiveness vision for the Finnish health sector [91] and the Sotedigi toolkit produced in collaboration with the business sector organizations [92].

Applying synthetic data in healthcare needs further ethical discussions and research to establish shared guidelines and best practices among public and private stakeholders, including dialogue with citizens. Public sector organizations such as Research Data Scotland (United Kingdom and Northern Ireland), South Australian Health (SA Health, Australia), the Personal Data Protection Commission (PDPC, Singapore), the United Nations Economic Commission for Europe (UNECE) and the Financial Conduct Authority (FCA, United Kingdom and Northern Ireland) have taken steps to explore the potential of synthetic data, either by launching strategic initiatives, setting up expert groups, or producing guidelines. In Germany, a National Data Infrastructure project known as NFDI4Health developed a SYNDAT platform for synthetic data [93]. Finland also has the expertise, infrastructure, and data reserves to explore the opportunities on a national level. It is important to recognize that synthetic data generation offers limited options for linking data from different sources, which is why the future EHDS capabilities will play a significant role in determining the collaborative potential.

Acknowledgments

This work has been carried out as part of the PRIVASA joint action (2021–2024) funded by Business Finland. The associated grant numbers are 37428/31/2020 for the University of Turku, 33961/31/2020 for the Turku University of Applied Sciences and 43450/31/2020 for the VTT Technical Research Centre of Finland. The authors wish to thank all consortium partners for their valuable contributions that have shaped the project's public research outcomes.

Conflict of interest

All authors declare no conflicts of interest regarding the works presented in this article.

References

1. V. Aula, Institutions, infrastructures, and data friction—reforming secondary use of health data in Finland, *Big Data Soc.*, **6** (2019), 1–13. http://dx.doi.org/10.1177/2053951719875980
2. *European commission, Proposal for a regulation of the European parliament and of the council on the European health data space*, European parliament, 2022. Available from: `https://www.europarl.europa.eu/legislative-train/theme-promoting-our-european-way-of-life/file-european-health-data-space`.
3. R. Lun, D. Siegal, T. Ramsay, G. Stotts, D. Dowlatshahi, Synthetic data in cancer and cerebrovascular disease research: a novel approach to big data, *PLoS ONE*, **19** (2024), e0295921. http://dx.doi.org/10.1371/journal.pone.0295921
4. E. Sizikova, A. Badal, J. G. Delfino, M. Lago, B. Nelson, N. Saharkhiz, et al., Synthetic data in radiological imaging: current state and future outlook, *Artif. Intell.*, **1** (2024), ubae007. http://dx.doi.org/10.1093/bjrai/ubae007
5. J. A. Thomas, R. E. Foraker, N. Zamstein, J. D. Morrow, P. R. Payne, A. B. Wilcox, Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: results from analyzing > 1.8 million SARS-CoV-2 tests in the United States national COVID cohort collaborative (N3C), *J. Am. Med. Inform. Asso.*, **29** (2022), 1350–1365. http://dx.doi.org/10.1093/jamia/ocac045
6. H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, A. Bano, Synthetic data generation: state of the art in health care domain, *Comput. Sci. Rev.*, **48** (2023), 100546. http://dx.doi.org/10.1016/j.cosrev.2023.100546
7. A. Gonzales, G. Guruswamy, S. R. Smith, Synthetic data in health care: a narrative review, *PLOS Digit Health*, **2** (2023), e0000082. http://dx.doi.org/10.1371/journal.pdig.0000082
8. S. James, C. Harbron, J. Branson, M. Sundler, Synthetic data use: exploring use cases to optimise data utility, *Discov. Artif. Intell.*, **1** (2021), 15. http://dx.doi.org/10.1007/s44163-021-00016-y
9. V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, et al., Synthetic data generation methods in healthcare: a review on open-source tools and methods, *Comput. Struct. Biotec.*, **23** (2024), 2892–2910. http://dx.doi.org/10.1016/j.csbj.2024.07.005
10. C. A. F. López, A. Elbi, On the legal nature of synthetic data, *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.
11. M. S. Gal, O. Lynskey, Synthetic data: legal implications of the data-generation revolution, *Iowa L. Rev.*, **109** (2023), 1087. http://dx.doi.org/10.2139/ssrn.4414385
12. J. Drechsler, A. C. Haensch, 30 years of synthetic data, *Statist. Sci.*, **39** (2024), 221–242. http://dx.doi.org/10.1214/24-STS927
13. J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, et al., Synthetic data—what, why and how? arXiv: 2205.03257. http://dx.doi.org/10.48550/arXiv.2205.03257
14. T. E. Raghunathan, J. P. Reiter, D. B. Rubin, Multiple imputation for statistical disclosure limitation, *J. Off. Stat.*, **19** (2003), 1.

15. K. El Emam, L. Mosquera, J. Bass, Evaluating identity disclosure risk in fully synthetic health data: model development and validation, *J. Med. Internet Res.*, **22** (2020), 23139. http://dx.doi.org/10.2196/23139

16. J. P. Reiter, Inference for partially synthetic, public use microdata sets, *Surv. Methodol.*, **29** (2003), 181–188.

17. H. Surendra, H. Mohan, A review of synthetic data generation methods for privacy preserving data publishing, *International Journal of Scientific and Technology Research*, **6** (2017), 95–101.

18. S. Mohiuddin, R. Gardiner, M. Crofts, P. Muir, J. Steer, J. Turner, et al., Modelling patient flows and resource use within a sexual health clinic through discrete event simulation to inform service redesign, *BMJ Open*, **10** (2020), e037084. http://dx.doi.org/10.1136/bmjopen-2020-037084

19. A. A. Tako, K. Kotiadis, C. Vasilakis, A. Miras, C. W. le Roux, Improving patient waiting times: a simulation study of an obesity care service, *BMJ Qual. Saf.*, **23** (2014), 373–381. http://dx.doi.org/10.1136/bmjqs-2013-002107

20. J. Yoon, M. Mizrahi, N. F. Ghalaty, T. Jarvinen, A. S. Ravi, P. Brune, et al., EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records, *NPJ Digit. Med.*, **6** (2023), 141. http://dx.doi.org/10.1038/s41746-023-00888-7

21. L. Juwara, A. El-Hussuna, K. El Emam, An evaluation of synthetic data augmentation for mitigating covariate bias in health data, *Patterns*, **5** (2024), 100946. http://dx.doi.org/10.1016/j.patter.2024.100946

22. S. Kaji, S. Kida, Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging, *Radiol. Phys. Technol.*, **12** (2019), 235–248. http://dx.doi.org/10.1007/s12194-019-00520-y

23. S. Dayarathna, K. T. Islam, S. Uribe, G. Yang, M. Hayat, Z. Chen, Deep learning based synthesis of MRI, CT and PET: review and analysis, *Med. Image Anal.*, **92** (2024), 103046. http://dx.doi.org/10.1016/j.media.2023.103046

24. K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, et al., MedGAN: medical image translation using GANs, *Comput. Med. Imag. Grap.*, **79** (2020), 101684. http://dx.doi.org/10.1016/j.compmedimag.2019.101684

25. J. Zhang, X. He, L. Qing, F. Gao, B. Wang, BPGAN: brain PET synthesis from MRI using generative adversarial network for multi-modal Alzheimer's disease diagnosis, *Comput. Meth. Prog. Bio.*, **217** (2022), 106676. http://dx.doi.org/10.1016/j.cmpb.2022.106676

26. M. J. Tadi, J. Teuho, R. Klén, E. Lehtonen, A. Saraste, C. S. Levin, Synthetic full dose cardiac PET images from low dose scans using conditional GANs, *Proceedings of IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2022, 1–2. http://dx.doi.org/10.1109/NSS/MIC44845.2022.10399148

27. D. Doncenco, Exploring medical image data augmentation and synthesis using conditional generative adversarial networks, B.S. Thesis, Turku University of Applied Sciences, 2022.

28. J. T. Huhtanen, M. Nyman, D. Doncenco, M. Hamedian, D. Kawalya, L. Salminen, et al., Deep learning accurately classifies elbow joint effusion in adult and pediatric radiographs, *Sci. Rep.*, **12** (2022), 11803. http://dx.doi.org/10.1038/s41598-022-16154-x

29. P. Movahedi, V. Nieminen, I. M. Perez, H. Daafane, D. Sukhwal, T. Pahikkala et al., Benchmarking evaluation protocols for classifiers trained on differentially private synthetic data, *IEEE Access*, **12** (2024), 118637–118648. http://dx.doi.org/10.1109/ACCESS.2024.3446913
30. A. R. Benaim, R. Almog, Y. Gorelik, I. Hochberg, L. Nassar, T. Mashiach, 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.*, **8** (2020), e16492. http://dx.doi.org/10.2196/16492
31. P. Movahedi, V. Nieminen, I. M. Perez, T. Pahikkala, A. Airola, Evaluating classifiers trained on differentially private synthetic health data, *Proceedings of IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, 2023, 748–753. http://dx.doi.org/10.1109/CBMS58004.2023.00313
32. B. Nowok, G. M. Raab, C. Dibben, Synthpop: bespoke creation of synthetic data in R, *J. Stat. Softw.*, **74** (2016), 1–26. http://dx.doi.org/10.18637/jss.v074.i11
33. A. Montanez, SDV: an open source library for synthetic data generation, Ph.D Thesis, Massachusetts Institute of Technology, 2018.
34. T. Li, N. Li, On the tradeoff between privacy and utility in data publishing, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, 517–526. http://dx.doi.org/10.1145/1557019.1557079
35. A. Slavković, J. Seeman, Statistical data privacy: a song of privacy and utility, *Annu. Rev. Stat. Appl.*, **10** (2023), 189–218. http://dx.doi.org/10.1146/annurev-statistics-033121-112921
36. B. Zhao, M. A. Kaafar, N. Kourtellis, Not one but many tradeoffs: privacy vs. utility in differentially private machine learning, *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2020, 15–26. http://dx.doi.org/10.1145/3411495.3421352
37. M. Hittmeir, R. Mayer, A. Ekelhart, A baseline for attribute disclosure risk in synthetic data, *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020, 133–143. http://dx.doi.org/10.1145/3374664.3375722
38. K. El Emam, L. Mosquera, X. Fang, Validating a membership disclosure metric for synthetic health data, *JAMIA Open*, **5** (2022), ooac083. http://dx.doi.org/10.1093/jamiaopen/ooac083
39. L. Sweeney, Simple demographics often identify people uniquely, Data Privacy Working Paper, 2000.
40. L. Sweeney, k-anonymity: a model for protecting privacy, *Int. J. Uncertain. Fuzz.*, **10** (2002), 557–570. http://dx.doi.org/10.1142/S0218488502001648
41. N. Li, T. Li, S. Venkatasubramanian, t-closeness: privacy beyond k-anonymity and l-diversity, *Proceedings of the 23rd international conference on data engineering*, 2007, 106–115. http://dx.doi.org/10.1109/ICDE.2007.367856
42. C. Dwork, A. Roth, The algorithmic foundations of differential privacy, *Found. Trends Theor. C.*, **9** (2014), 211–407. http://dx.doi.org/10.1561/0400000042
43. M. Finck, F. Pallas, They who must not be identified—distinguishing personal from non-personal data under the GDPR, *Int. Data Priv. Law*, **10** (2020), 11–36. http://dx.doi.org/10.1093/idpl/ipz026

44. A. Cohen, K. Nissim, Towards formalizing the GDPR's notion of singling out, *PNAS*, **117** (2020), 8344–8352. http://dx.doi.org/10.1073/pnas.1914598117

45. M. Veale, R. Binns, L. Edwards, Algorithms that remember: model inversion attacks and data protection law, *Phil. Trans. R. Soc. A*, **376** (2018), 20180083. http://dx.doi.org/10.1098/rsta.2018.0083

46. C. Sun, J. van Soest, M. Dumontier, Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy, *J. Biomed. Inform.*, **143** (2023), 104404. http://dx.doi.org/10.1016/j.jbi.2023.104404

47. J. Jordon, J. Yoon, M. van der Schaar, PATE-GAN: generating synthetic data with differential privacy guarantees, *Proceedings of International Conference on Learning Representations*, 2019, 1–29.

48. N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, L. Sweeney, Privacy preserving synthetic data release using deep learning, In: *Machine learning and knowledge discovery in databases*, Cham: Springer, 2019, 510–526. http://dx.doi.org/10.1007/978-3-030-10925-7_31

49. I. Montoya Perez, P. Movahedi, V. Nieminen, A. Airola, T. Pahikkala, Does differentially private synthetic data lead to synthetic discoveries? *Methods Inf. Med.*, in press. http://dx.doi.org/10.1055/a-2385-1355

50. M. I. Khan, M. A. Azeem, E. Alhoniemi, E. Kontio, S. A. Khan, M. Jafaritadi, Regularized weight aggregation in networked federated learning for glioblastoma segmentation, In: *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries*, Cham: Springer, 2022, 121–132. http://dx.doi.org/10.1007/978-3-031-44153-0_12

51. M. I. Khan, M. Jafaritadi, E. Alhoniemi, E. Kontio, S. A. Khan, Adaptive weight aggregation in federated learning for brain tumor segmentation, In: *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries*, Cham: Springer, 2022, 455–469. http://dx.doi.org/10.1007/978-3-031-09002-8_40

52. M. I. Khan, E. Alhoniemi, E. Kontio, S. A. Khan, M. Jafaritadi, RegAgg: a scalable approach for efficient weight aggregation in federated lesion segmentation of brain MRIs, *Proceedings of Eighth International Conference on Fog and Mobile Edge Computing (FMEC)*, 2023, 101–106. http://dx.doi.org/10.1109/FMEC59375.2023.10306171

53. J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, F. Wang, Federated learning for healthcare informatics, *J. Healthc. Inform. Res.*, **5** (2021), 1–19. http://dx.doi.org/10.1007/s41666-020-00082-4

54. M. Giuffrè, D. L. Shung, Harnessing the power of synthetic data in healthcare: innovation, application, and privacy, *NPJ Digit. Med.*, **6** (2023), 186. http://dx.doi.org/10.1038/s41746-023-00927-3

55. D. Shanley, J. Hogenboom, F. Lysen, L. Wee, A. Lobo Gomes, A. Dekker, et al., Getting real about synthetic data ethics: are AI ethics principles a good starting point for synthetic data ethics? *EMBO Rep.*, **25** (2024), 2152–2155. http://dx.doi.org/10.1038/s44319-024-00101-0

56. B. N. Jacobsen, Machine learning and the politics of synthetic data, *Big Data Soc.*, **10** (2023), 1–12. http://dx.doi.org/10.1177/20539517221145372

57. C. D. Whitney, J. Norman, Real risks of fake data: synthetic data, diversity-washing and consent circumvention, *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, 1733–1744. http://dx.doi.org/10.1145/3630106.3659002

58. G. Ganev, B. Oprisanu, E. De Cristofaro, Robin Hood and Matthew effects: differential privacy has disparate impact on synthetic data, *Proceedings of the 39th International Conference on Machine Learning*, 2022, 6944–6959.

59. T. Hayashi, D. Cimr, H. Fujita, R. Cimler, Interpretable synthetic signals for explainable one-class time-series classification, *Eng. Appl. Artif. Intell.*, **131** (2024), 107716. http://dx.doi.org/10.1016/j.engappai.2023.107716

60. J. Vaiste, Ethical implications of AI-generated synthetic health data, HAL Id: hal-04216538.

61. J. S. Franklin, K. Bhanot, M. Ghalwash, K. P. Bennett, J. McCusker, D. L. McGuinness, An ontology for fairness metrics, *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, 265–275. http://dx.doi.org/10.1145/3514094.3534137

62. K. Bhanot, M. Qi, J. S. Erickson, I. Guyon, K. P. Bennett, The problem of fairness in synthetic healthcare data, *Entropy*, **23** (2021), 1165. http://dx.doi.org/10.3390/e23091165

63. T. Farrand, F. Mireshghallah, S. Singh, A. Trask, Neither private nor fair: impact of data imbalance on utility and fairness in differential privacy, *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*, 2020, 15–19. http://dx.doi.org/10.1145/3411501.3419419

64. V. Volovici, N. L. Syn, A. Ercole, J. J. Zhao, N. Liu, Steps to avoid overuse and misuse of machine learning in clinical research, *Nat. Med.*, **28** (2022), 1996–1999. http://dx.doi.org/10.1038/s41591-022-01961-6

65. A. S. Hashemi, A. Soliman, J. Lundström, K. Etminani, Domain knowledge-driven generation of synthetic healthcare data, *Stud. Health Technol. Inform.*, **302** (2023), 352–353. http://dx.doi.org/10.3233/SHTI230136

66. J. Latner, M. Neunhoeffer, J. Drechsler, Generating synthetic data is complicated: know your data and know your generator, In: *Privacy in statistical databases*, Cham: Springer, 2024, 115–128. http://dx.doi.org/10.1007/978-3-031-69651-0_8

67. F. K. Dankar, M. K. Ibrahim, L. Ismail, A multi-dimensional evaluation of synthetic data generators, *IEEE Access*, **10** (2022), 11147–11158. http://dx.doi.org/10.1109/ACCESS.2022.3144765

68. M. Miletic, M. Sariyar, Assessing the potentials of LLMs and GANs as state-of-the-art tabular synthetic data generation methods, In: *Privacy in statistical databases*, Cham: Springer, 2024, 374–389. http://dx.doi.org/10.1007/978-3-031-69651-0_25

69. R. Hamon, H. Junklewitz, I. Sanchez, *Robustness and explainability of artificial intelligence*, Luxembourg: Publications Office of the European Union, 2020. http://dx.doi.org/10.2760/57493

70. M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, D. Rankin, Synthetic data generation for tabular health records: a systematic review, *Neurocomputing*, **493** (2022), 28–45. http://dx.doi.org/10.1016/j.neucom.2022.04.053

71. K. Perkonoja, K. Auranen, J. Virta, Methods for generating and evaluating synthetic longitudinal patient data: a systematic review, arXiv: 2309.12380. http://dx.doi.org/10.48550/arXiv.2309.12380

72. J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, X. Xiao, PrivBayes: private data release via Bayesian networks, *ACM T. Database Syst.*, **42** (2017), 25. http://dx.doi.org/10.1145/3134428

73. J. de Benedetti, N. Oues, Z. Wang, P. Myles, A. Tucker, Practical lessons from generating synthetic healthcare data with Bayesian networks, In: *ECML PKDD 2020 workshops*, Cham: Springer, 2020, 38–47. http://dx.doi.org/10.1007/978-3-030-65965-3_3

74. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial nets, *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014, 2672–2680.

75. D. P. Kingma, M. Welling, Auto-encoding variational Bayes, arXiv: 1312.6114. http://dx.doi.org/10.48550/arXiv.1312.6114

76. C. Yan, Y. Yan, Z. Wan, Z. Zhang, L. Omberg, J. Guinney, et al., A multifaceted benchmarking of synthetic electronic health record generation models, *Nat. Commun.*, **13** (2022), 7609. http://dx.doi.org/10.1038/s41467-022-35295-1

77. S. Biswal, S. Ghosh, J. Duke, B. Malin, W. Stewart, C. Xiao, J. Sun, EVA: generating longitudinal electronic health records using conditional variational autoencoders, *Proceedings of the 6th Machine Learning for Healthcare Conference*, 2021, 260–282.

78. F. K. Dankar, M. Ibrahim, Fake it till you make it: guidelines for effective synthetic data generation, *Appl. Sci.*, **11** (2021), 2158. http://dx.doi.org/10.3390/app11052158

79. C. Yan, Z. Zhang, S. Nyemba, Z. Li, Generating synthetic electronic health record data using generative adversarial networks: tutorial, *JMIR AI*, **3** (2024), e52615. http://dx.doi.org/10.2196/52615

80. V. Nieminen, T. Pahikkala, A. Airola, Empirical evaluation of amplifying privacy by subsampling for GANs to create differentially private synthetic tabular data, *Proceedings of TKTP 2023: Annual Symposium for Computer Science*, 2023, 72–81.

81. A. Alaa, B. van Breugel, E. S. Saveliev, M. van der Schaar, How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models, *Proceedings of the 39th International Conference on Machine Learning*, 2022, 290–306.

82. A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, K. P. Bennett, Generation and evaluation of privacy preserving synthetic health data, *Neurocomputing*, **416** (2020), 244–255. http://dx.doi.org/10.1016/j.neucom.2019.12.136

83. J. Yoon, L. N. Drumright, M. van der Schaar, Anonymization through data synthesis using generative adversarial networks (ADS-GAN), *IEEE J. Biomed. Health*, **24** (2020), 2378–2388. http://dx.doi.org/10.1109/JBHI.2020.2980262

84. V. B. Vallevik, A. Babic, S. E. Marshall, E. Severin, H. M. Brøgger, S. Alagaratnam, et al., Can I trust my fake data—a comprehensive quality assessment framework for synthetic tabular data in healthcare, *Int. J. Med. Inform.*, **185** (2024), 105413. http://dx.doi.org/10.1016/j.ijmedinf.2024.105413

85. Z. Azizi, S. Lindner, Y. Shiba, V. Raparelli, C. M. Norris, K. Kublickiene, et al., A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health, *Sci. Rep.*, **13** (2023), 11540. http://dx.doi.org/10.1038/s41598-023-38457-3

86. M. Hernandez, G. Epelde, A. Beristain, R. Álvarez, C. Molina, X. Larrea, et al., Incorporation of synthetic data generation techniques within a controlled data processing workflow in the health and wellbeing domain, *Electronics*, **11** (2022), 812. http://dx.doi.org/10.3390/electronics11050812

87. C. Little, M. Elliot, R. Allmendinger, Federated learning for generating synthetic data: a scoping review, *Int. J. Popul. Data Sci.*, **8** (2023), 2158. http://dx.doi.org/10.23889/ijpds.v8i1.2158

88. J. W. Kim, B. Jang, Privacy-preserving generation and publication of synthetic trajectory microdata: a comprehensive survey, *J. Netw. Comput. Appl.*, **230** (2024), 103951. http://dx.doi.org/10.1016/j.jnca.2024.103951

89. C. Alloza, B. Knox, H. Raad, M. Aguilà, C. Coakley, Z. Mohrova, et al., A case for synthetic data in regulatory decision-making in Europe, *Clin. Pharmacol. Ther.*, **114** (2023), 795–801. http://dx.doi.org/10.1002/cpt.3001

90. A. Beduschi, Synthetic data protection: towards a paradigm change in data regulation? *Big Data Soc.*, **11** (2024), 1–5. http://dx.doi.org/10.1177/20539517241231277

91. P. Lehto, S. Malkamäki, *The Finnish health sector growth and competitiveness vision 2030*, Helsinki: Sitra, 2023.

92. *Finnish association of private care providers, Sotedigin työkalupakista eväitä tiedon hyödyntämiseen sote-palveluissa*, Hyvinvointiala Hali ry, 2023. Available from: `https://www.hyvinvointiala.fi/sotedigin-tyokalupakista-evaita-tiedon-hyodyntamiseen-sote-palveluissa/`.

93. S. Moazemi, T. Adams, H. G. NG, L. Kühnel, J. Schneider, A. F. Näher, et al., NFDI4Health workflow and service for synthetic data generation, assessment and risk management, *Stud. Health Technol. Inform.*, **317** (2024), 21–29. http://dx.doi.org/10.3233/SHTI240834

AIMS Press

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)