Computer Science Review 48 (2023) 100546

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

Synthetic data generation: State of the art in health care domain ☆

Hajra Murtaza [a,*], Musharif Ahmed [a], Naurin Farooq Khan [a], Ghulam Murtaza [b], Saad Zafar [a], Ambreen Bano [c]

[a] Faculty of Computing, Riphah International University, I-14, Hajj Complex, Islamabad, Pakistan
[b] College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Islamabad, Pakistan
[c] Department of Mathematics and Statistics, Riphah International University, I-14, Hajj Complex, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 1 September 2022
Received in revised form 20 January 2023
Accepted 12 February 2023
Available online 26 February 2023

Keywords:
Synthetic data
Health informatics
Data privacy
Privacy preserving data publishing
Medical informatics
Generative adversarial networks
Electronic health records

ABSTRACT

Recent progress in artificial intelligence and machine learning has led to the growth of research in every aspect of life including the health care domain. However, privacy risks and legislations hinder the availability of patient data to researchers. Synthetic data (SD) has been regarded as a privacy-safe alternative to real data and has lately been employed in many research and academic endeavors. This growing body of research needs to be consolidated for the researchers and practitioners to gain a quick and fruitful comprehension of the state of the art in synthetic data generation in health care. The purpose of this study is to collate and synthesize the current state of synthetic data generation following a narrative review of 70 peer-reviewed studies discussing privacy-preserving synthetic medical data generation techniques. The literature shows the effectiveness of synthetic datasets for different applications in research, academics, and testing according to existing statistical and task-based utility metrics. However, the focus on longitudinal synthetic data seems deficient. Moreover, a unified metric for generic quality assessment of synthetic data is lacking. The results of this review will serve as a quick reference guide for the researchers and practitioners in the healthcare domain to select a suitable synthetic data strategy for their application based on its strengths and weaknesses and pave the path for further research and development in healthcare.

Contents

☆ Funding: No funding was received to assist with the preparation of this manuscript.
* Corresponding author.
E-mail addresses: hajra.murtaza@riphah.edu.pk (H. Murtaza), musharif.ahmed@riphah.edu.pk (M. Ahmed), naurin.zamir@riphah.edu.pk (N.F. Khan), gmurtaza.ceme@ceme.nust.edu.pk (G. Murtaza), saad.zafar@riphah.edu.pk (S. Zafar), ambreen.bano@riphah.edu.pk (A. Bano).

https://doi.org/10.1016/j.cosrev.2023.100546
1574-0137/

1. Introduction

Information and Communication Technologies (ICT) have impacted every aspect of life. Digitization of the hospital environment is a revolutionary shift from traditional human-centric practice to advanced technology-assisted medical care. The integration of ICT in the medical domain can be seen in routine patient care through advanced telemedicine, Electronic Health Records (EHR) in complex medical informatics, personal health assistants, and medical decision support systems. This integration has garnered more pronounced benefits compared to other sectors through improvement in the efficiency, pervasiveness, accuracy, and reliability of health services [1]. Fast-paced technological developments have resulted in the adoption of the "new" with quick obsolescence of the "old" [2] that hinges on the availability of the dataset. High-quality healthcare data is essential for high-quality research, better development initiatives, and outcomes, informed medical decisions, and better quality of life [3]. Quality assurance of digital health systems requires large datasets to test the application/interventions in realistic clinical scenarios for clinical validation. Unfortunately, the availability of health care data poses some unique challenges [4], prohibiting open sharing primarily due to privacy concerns [5,6]. Medical datasets usually contain sensitive personal information such as diagnoses, treatments, and billing records. Exposing this information is undesirable and raises ethical, financial, and legal issues [7]. Therefore, the public release of medical data is subject to restrictions due to stringent privacy regulations such as Health Insurance Portability Accountability Act (HIPAA) [8], and General Data Protection Regulation (GDPR) [9]. These restrictions render the wealth of medical data impotent for further medical informatics research [5]. To overcome these restrictions, researchers are finding ways to enable the privacy-safe dissemination of health data to the research community. Earlier initiatives involved data masking and anonymization techniques to transform the data in a way that preserved its statistical properties while blocking any privacy leakages [10]. However, the resulting data loses much of its truthfulness and offers little use for advanced research. Moreover, anonymized data is susceptible to residual privacy risks and thus fails to fulfill the fundamental requirement [11].

Researchers have proposed synthetic data (SD) as an alternative to data transformation. Synthetic data is created artificially, possesses the same statistical characteristics as the original, and yet shows better resilience to privacy attacks [12]. SD is required to exhibit the same distribution as well as correlational structure as the original data also known as "realism" or "resemblance". Realistic synthetic data can replace the original data in many applications [12] and can be of utmost significance for medical research where real datasets are unavailable. Synthetic data is being used for testing and evaluation [13,14], and statistical disclosure control [15]. However, generating synthetic medical data is surrounded by unique challenges because of its inherent complexity and longitudinal nature [16]. There has been a sharp rise in research publications in the field of synthetic medical data generation in the past few years and wider adoption of SD is expected in the future [17]. Researchers are experimenting with different methods to generate realistic synthetic data and have proposed many different quality evaluation metrics to assess its plausibility as a substitute for real data. This has introduced an array of new concepts, terms, techniques, and metrics in the literature. There is a need to consolidate this body of knowledge for better understanding.

List of Acronyms

AI	Artificial Intelligence
AD	Attribute Disclosure
AE	Autoencoders
BN	Bayesian Networks
CPGs	Clinical Procedural Guidelines
CNN	Convolutional Neural Network
DD	Data Driven [Approaches]
DUA	Data Use Agreement
DT	Decision Tree
DL	Deep Learning
DP	Differential Privacy
DWP	Dimension Wise Prediction
DWS	Dimension Wise Statistics
DPM	Disease Progression Models
DM	Domain Model
ECG	Electrocardiogram
EEG	Electroencephalogram
EMR	Electronic Medical Records
GAN	Generative Adversarial Networks
GM	Generative Models
HIS	Health Incidence Statistics
HBD	Hybrid [Approaches]
KDE	Kernel Density Estimation
KD	Knowledge Driven [Approaches]
KLD	Kullback Leibler Divergence
LSTM	Long Short-Term Memory
ML	Machine Learning
MLE	Maximum Likelihood Estimates
MI	Membership Inference
MC	Monte Carlo
MLP	Multi-Layer Perceptron
NDMS	Non-metric Multidimensional Scaling
PHI	Personal Health Indicators
RDT	Randomized Decision Tree
RNN	Recurrent Neural Network
SDC	Statistical Disclosure Control
SD	Synthetic Data
STM	State Transition Machine
SDG	Synthetic Data Generation
VAE	Variational Autoencoder
WD	Wasserstein Distance

The rest of the paper is organized as follows. In Section 2, we present a brief background of the topic to introduce the important domain concepts followed by related work in Section 3 and research methodology in Section 4. We present a narrative review of the synthetic medical data generation techniques and SD evaluation in Section 5 and the limitations, conclusion, and future directions in Sections 6, 7, and 8 respectively.

2. Synthetic medical data generation — Background

2.1. Significance, challenges, and applications

Before putting it into practice, every new solution, especially for medical practitioners and researchers must be thoroughly validated for being realistic and effective [18],. This needs a vast amount of healthcare data. Large masses of medical data are produced at health delivery sites but the same is not available for secondary use due to privacy laws [5]. Anonymization techniques fail to provide a comprehensive privacy solution for high-dimensional health data [19] thus shifting attention to synthetic data as a privacy-preserving data-sharing method [12]. Synthetic datasets offer several benefits [20]:

a. Resilience to privacy attacks [12]
b. An inexpensive and convenient alternative to real data [13,21]
c. Specific instances as per requirements can be produced [22]
d. Virtually unlimited supply of real-like data
e. Lesser regulatory restrictions, faster dissemination, and thus shorter times to-insight [23]

Synthetic data has been employed in the following health-related scenarios.

- **Forecasting and Planning**: Reliable health forecasts facilitate better healthcare planning and conduct as well as lifestyle planning. Buczak et al. [24] suggest using synthetic data to train models that could predict the onset of an epidemic and alert the authorities to take necessary action well in time. Similarly, Liu et al. suggested using synthetic medical data to support urban planning for an improved lifestyle [25]. Foraker et al. have demonstrated the utility of SD for geospatial data analysis in [23]
- **Design and Evaluation of New Health Technology and Algorithms**: Multiple researchers have proposed to use realistic synthetic data as a substitute for actual data for testing any new healthcare devices and algorithms [17,21,26–28]. For example, diabetic foot data was generated [27] to assist in the development of the diabetic foot treatment insole [29]. Chen et al. [17] note the utility of synthetic data to test new algorithms for the Artificial Intelligence – Software As Medical Device (AI-SaMD) initiative [30].
- **First Order Feasibility:** In some cases, de-identified real data is available, but acquiring it involves high costs and intricate legal requirements. Institutions can employ synthetic data to make a first-order cheaper assessment of the approximate utility of the real data for the task under consideration before actual procurement [31]. This first-level analysis of utility can save cost and effort.
- **Data Science Competitions and Hackathons**: Hackathons and competitions are a quick way of getting innovative solutions to technical problems. However, the sensitivity of health data hinders using such platforms in medical science. Synthetic Corona Virus data has been employed in hackathons and conferences without any privacy risks [32].
- **Academic Settings:** Hands-on practical training in the classroom has been linked with increased understanding and longer retention of the concepts [33]. Researchers report promising outcomes with synthetic data use for a classroom data science challenge [28,34].
- **Testing and Benchmarking:** Synthetic data have long been used to generate test data for software and hardware tools in various fields and recently have been employed in the medical domain as well [21,28,35,36]. More recently, Emam et al. presented an interesting application of SD to obtain closer estimates of re-identification risks associated with anonymized samples of data [37].
- **Data Augmentation:** Medical decision-making can benefit from advanced AI techniques but need highly accurate models to minimize the risk of inaccurate outcomes which may prove fatal. Training of such high-performing AI agents requires large volumes of quality data which are hard to

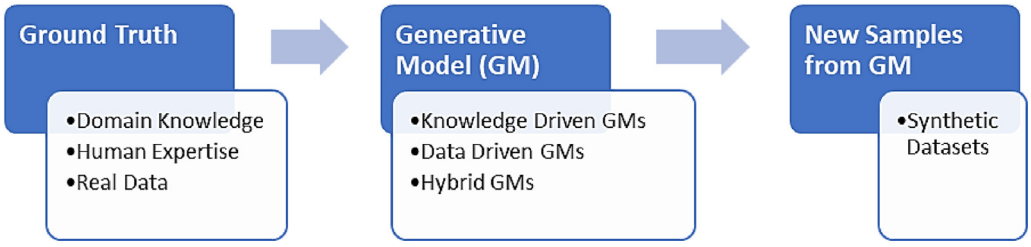

Fig. 1. A generic SDG process.

Table 1
Comparison of KD, DD, and HBD approaches for synthetic data generation.

SDG approach	Knowledge driven	Data driven	Hybrid
Source of ground truth	Ground truth is derived from public knowledge i.e., theory and experts, and built into a hand-crated generative model for SD generation	Derives the ground truth from real data through an automated model estimation mechanism	The generative model is built with ground truth which is an amalgamation of theory and real data.
Pros	• No dependence on hard-to-acquire real data. • Abundant public knowledge is available. • Provide complete flexibility and control over the generative process. • GM can be manipulated at a micro level to fine-tune the output. • Face no disclosure risk.	• DD approaches are dependent on the quality and quantity of available real data. • Real datasets are few and do not always provide the required level of detail. • DD approaches are fully automated and provide a mechanism to exercise some control over the generative process (e.g., conditional generation).	• Combines the benefits of both KD and DD methods. • The lack of sufficient real data can be supplemented with freely available domain knowledge. • Automation from DD and control from KD allow for sophisticated generative modeling. • Lower disclosure risk compared to fully DD generation.
Cons	• Model Crafting is arduous and laborious. • Susceptible to omissions and inconsistencies. • The quality of SD is highly sensitive to the skill of the expert constructing the model. • Oblivious to the ground truth hidden in real field data not known in the theory yet.	• Real Data Acquisition is a big challenge. • Biases of the real data may seep into the SD. • The quality of SD is sensitive to the quality and quantity of available real data. • Susceptible to membership and attribute disclosure attacks.	• Face the data acquisition problems as the DD approaches. • The KD part might not always be automatable. • Deciding the appropriate proportions of DD and KD elements for generative modeling is crucial to cut an optimal balance between realism and privacy.

get hold of. Whenever available, the useable data is insufficient for data mining and machine learning tasks with such tight performance requirements. Many researchers [38–43] have shown the improved performance of models when trained with augmented datasets containing a mix of real and synthetic records.

Synthetic data needs to be sufficiently "realistic" to replace real data in practical applications. That means it must exhibit the same structural and statistical characteristics. Synthetic data generation essentially estimates a structural model extracted from the "ground truth" in the real data. Health data contains high-dimensional records with heterogeneous features, recorded repeatedly over time [3]. This longitudinal nature of medical data introduces complex sequential and temporal inter-dependencies between the features, thus posing unique modeling challenges [16]. Synthetic data generation consists of three fundamental steps as shown in Fig. 1. (a) Collecting relevant facts (data) and theories (knowledge) from the real-world, (b) Building the generative model which encompasses the necessary ground truth, and (c) Extracting new unseen samples from the model. We use the term "Generative Model" (GM) to refer to a structured representation of the relevant ground truth which must be incident in the synthetic samples. The model itself may depict a discriminative (e.g. decision tree) or generative (e.g. GAN) structure, as per the communal understanding [44]. Synthetic data can be produced at a fast pace, low cost, and large volumes and can be made to have desired characteristics. However, the true potential of synthetic data is yet to be exploited in the medical domain [45].

2.2. Synthetic data generation approaches

Synthetic data generation methods are generally classified into three categories based on the source of ground truth [46]. These include (1) Knowledge-Driven Methods, (2) Data-Driven Methods and (3) Hybrid Methods. We present a brief introduction to these approaches in the following subsection. Table 1 presents a comparison of these approaches.

2.2.1. Knowledge-Driven (KD) approaches

Also known as *Theory driven,* these approaches derive the ground truth from the publicly available domain-specific knowledge which can be derived from academic or research documents, web resources, and human expertise to name a few. These techniques require manual curation of the generative model in form of rules, statistical, mathematical, or computational models, or computer programs. Real data are rarely involved in this process making these methods resilient to disclosure attacks. Nevertheless, the utility of the KD synthetic data greatly depends upon the ability, knowledge, and skill of the model curator. A good generative model should embody maximum coverage and completeness of the ground truth. However, complete coverage is hard to achieve because of the high dimensionality and complex correlations in medical data. Moreover, errors and omissions due to human intervention can sabotage the truthfulness of the generated data. Another important aspect is the oblivion of domain theory towards the "positive deviances" inevitably happening in medical practice [47]. Moreover, data from the field of practice contain hidden patterns not known to the world yet. An exclusively theory-based model may lack a significant portion of this

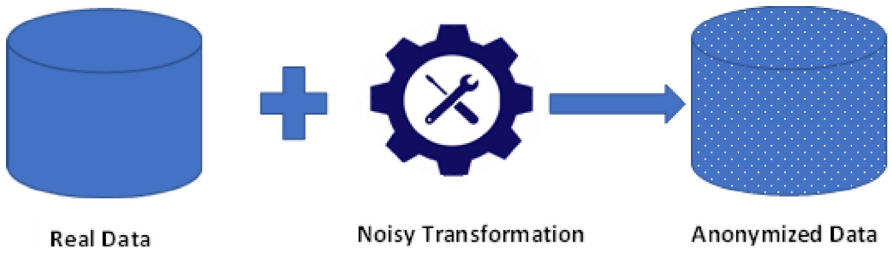

Fig. 2. Transformative DD-SDG.

ground truth and may lose real-life clinical relevance. Further, only some aspects of the KD based methods can be automated.

2.2.2. Data-driven approaches

Data-driven techniques derive the generative model from the actual data. Real data recorded by observing the true events, embodies an indirect representation of the domain theory [22] along with the "positive deviances" occurring in the field of practice [47]. Data-driven methods have been studied more rigorously than theory-based approaches mainly because real field data is better at capturing ground realities [22]. Secondly, these approaches can be fully automated, resulting in better efficiency. Nevertheless, the quality of generated data largely depends upon the quality of original data in terms of correctness, diversity, and coverage. Moreover, these approaches face the bootstrapping problem i.e., need real data to solve the lack of real data. We classify the data-driven approaches as working in transformative or simulation modes.

- *Transformative Methods:* These methods transform the original data using different masking operations (Fig. 2), with varying degrees of noise, before publishing, to guard against privacy leakage. The process is similar to the classical Statistical Disclosure Control (SDC) techniques [48]. Popular transformative approaches include micro-aggregation, multiple imputations, perturbation, and data anonymization. These approaches must balance a tough tradeoff between privacy and the utility of data. However, transformations are intrinsically lossy and may result in significant degradation of the data truthfulness to satisfy the high privacy requirements, especially in the case of medical data. This can result in the synthetic data losing clinical relevance [11]. The present study does not discuss this category.
- *Simulations:* In these techniques, the generative models estimate the probability distribution of the real data, and synthetic samples are drawn from it. This leverages semantic similarity in the underlying structure of synthetic and real data. Classical methods such as Maximum Likelihood Estimates (MLE) and Copulas along with more recent machine learning models such as VAEs and GANs have been employed by the researchers in this category. Since the synthetic samples are drawn from the same underlying distributions, they are expected to exhibit the same aggregate structure as real data but are sufficiently different on the micro level to avoid privacy leakages of the real values. Fig. 3 demonstrates a high-level representation of a typical simulation-style SDG. This article includes the simulation-based methods only.

2.2.3. Hybrid methods

Hybrid approaches derive the ground truth from both, theory, and real data. The generative model learns the truth from data and annotates it with "advice" from domain-theory/experts. This helps to control the induction of real-data biases into the generative model. Advice from the domain expert accelerates model convergence [49] and also facilitates steering the generation towards the desired region of the sample space. Data-driven generative models are limited by the quality and granularity of real data and knowledge-driven approaches by the capacity of human curators. Combining these methods in hybrid approaches can introduce more flexibility and comprehensiveness in the generative model. More research in this direction can enhance the current state of synthetic medical data generation.

2.3. Ground truth selection for synthetic data generation

Abundant public knowledge is available from a variety of sources, in a variety of different formats for knowledge-driven approaches. These include academic and research publications, web resources, local health departments' publications, and domain experts. But most of this information is not machine ready and needs a lot of manual effort and preprocessing to consolidate it into the generative model. This process is inherently slow and error-prone. Data-driven approaches, on the other hand, are completely automatable, resulting in better reliability and efficiency. However, the quality of synthetic data is tightly bounded by the quality of and the degree of truth in the input data. Medical data sets can be classified considering two perspectives; ***(1) Privacy Perspective*** which categorizes datasets as Private, Restricted, and Public data and **(2)** ***Structural Perspective*** which classifies the datasets into Snapshot/Aggregate, Longitudinal, or Timeseries datasets. *Private data* is the original data recorded at health care sites which usually contains sensitive information which is restricted from sharing by privacy laws. The health data owners can, however, create privacy-safe versions of restricted data through artificial synthesis or anonymization, which we call public data. The *public datasets* contain synthetic or anonymized versions of the real records to mitigate disclosure risk. Some datasets such as N3C [50] have *restricted* access granted only to the trusted parties under Data Use Agreements (DUA) to ensure the safe and sensible use of data. *Longitudinal data* contains patients' health information recorded at different points in time while the aggregate or snapshot datasets present just a cross-sectional view of the overall health condition of a patient. The EHRs recorded at health care sites are longitudinal observations containing information at a finer granularity [3]. Parts of this information are considered sensitive and privacy laws mandate anonymization of these before sharing [5]. For this reason, privacy-safe *snapshots or aggregated* representation of the longitudinal data is generally obtained for public sharing, which excludes identifying or other sensitive information. While fine-grained datasets offer maximum utility and potential for research and development, they are more susceptible to privacy leakages which may cause substantial financial, emotional, and social damages. This explains the high proportion of studies utilizing snapshot or aggregated real data representations (Fig. 6).

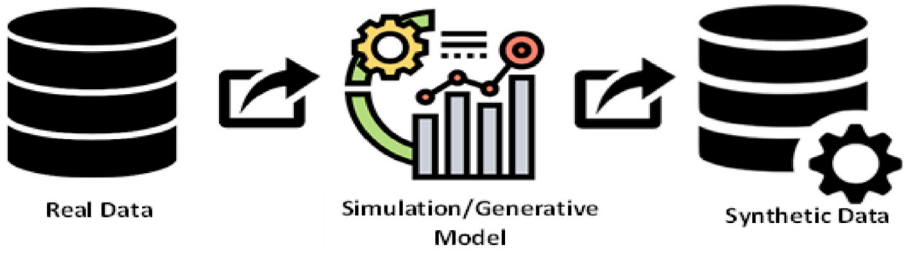

Fig. 3. Simulation-based DD SDG.

2.4. "Syntheticity" of the synthetic data

Synthetic data contains artificial values. We introduce the term "syntheticity" to refer to the proportion of artificial values in a dataset. Higher syntheticity means higher proportion of synthetic vs. real values in the data. The syntheticity of data bears an indirect relationship with privacy risks. Based on syntheticity, datasets can be either partially or fully synthetic as described in the following sub-sections.

2.4.1. Partially synthetic data

Partially or semi-synthetic datasets contain a mixture of real and artificial values, generally produced as a result of transformative SDG approaches. Modifications are made to the observed values of only certain sensitive parts of the dataset while non-sensitive values are left unchanged. Such datasets are desirable when only specific attributes need disclosure protection e.g., Individuals are comfortable sharing their educational information but not their financial status. It is computationally less expensive to generate partial synthetic data, yet it poses high disclosure risks because of the presence of original values. Also, the lossy transformations may render partially synthetic datasets unfit for advanced research [51]. Multiple Imputations is the most popular partially synthetic data generation technique.

2.4.2. Fully synthetic data

Fully synthetic data do not retain any of the original values. Realistic fully synthetic datasets offer high utility with minimal disclosure risks but are not completely impervious to privacy attacks. Researchers emphasize privacy evaluation of fully synthetic data generation techniques also. Data-driven synthetic data faces membership and attribute disclosure risks [52]. The sensitivity of health data and sharing restrictions make fully synthetic data generation a better candidate for this domain. This is also supported by the high proportion of fully synthetic medical data studies found in the literature as shown in Fig. 5.

2.5. Synthetic data - Quality attributes

Researchers regard synthetic data as a promising substitute for the original data provided it exhibits two quality attributes [28] (1) Realism and (2) Privacy Preservation. The following subsections briefly introduce the synthetic data quality parameters.

2.5.1. Realism

The quality of synthetic data being similar in behavior to the original data is termed as "Realism" [53] or "Resemblance" [28]. The higher the realism in synthetic data the better it can substitute the original, in practical settings. However, realism needs to be bounded against the privacy requirements; Synthetic records should resemble the real data but must be sufficiently different so as not to leak any sensitive information about the real data they are simulated from [28]. Realism validation assesses the plausibility of synthetic records in terms of the statistical (univariate and multivariate) similarity between the synthetic and real distributions as well as how well synthetic data can substitute real data for data mining and analysis applications [54]. In addition to the quantitative evaluation of these characteristics, researchers have emphasized the qualitative assessment of synthetic records by medical experts.

2.5.2. Privacy preservation

Privacy metrics measure the potential privacy disclosure risk. Knowledge-driven synthetic data is intrinsically privacy safe since no actual data is ever used in the generation process. Data-driven methods derive the generative knowledge from the real data and are susceptible to privacy leakages. Partially synthetic datasets, being similar to anonymized data, face the same privacy threats including identity, membership, and attribute disclosures [55]. Fully synthetic data, on the other hand, provide relatively higher privacy assurances but are not completely invulnerable [12] and may give in to membership and attribute inference attacks. Researchers further recognize that presence of rare instances may also result in meaningful identity disclosures [56].

3. Related work

The popularity of synthetic data is evident from the increasing number of publications every passing year manifested by the sharp upward trend in Fig. A.1 – Appendix A. To the best of our knowledge, this is the first systematic literature review of synthetic medical data generation comprehensively incorporating the different SDG approaches as well as multiple dataset formats. However, some reviews focusing on certain aspects of the SDG life cycle, or a specific SDG approach have recently been published, which we discuss below.

Hernandez et al. published an SLR on SDG for tabular healthcare data [57]. However, this work includes studies from the period 2016–2021, missing a significant proportion of studies older than the selected time frame. Moreover, the study [57] sought answers to research questions with a focus on GAN-based approaches for tabular data generation which are significantly different from the goals of the present study. The authors presented a narration of 34 studies and highlighted the various resemblance and utility metrics for the evaluation of SD employed by each. The authors also presented a subjective comparison of the included methods as excellent, good, or poor concerning resemblance, utility, and privacy. The assessment was based on the results reported by the respective studies that employed different metrics and thus the authors concluded that standardization of the metrics is an essential next step for a more objective assessment of the SDG methods. Our study takes one step further in this direction by presenting a classification of the various SD evaluation metrics used by the researchers so far. Moreover, we consider privacy as a core quality attribute for SD [58] and do not lose this focus in our discussion as opposed to [57]. Jordon et al. surveyed the generic synthetic data generation research in different domains and provided a high-level discussion of the

core concepts in [59]. A more detailed version of the survey was presented in [60], but, the study did not follow a systematic approach and missed some important research works. Nevertheless, the authors highlighted important directions for further research and the limitations of the existing methods. Another short study [20] reviewed the SDG with a focus on different use cases as a means to develop better SD solutions for the pharmaceutical domain. The study provided a high-level discussion of the utility and privacy evaluation metrics and highlighted a need for unified metrics for these to foster better trust in SD and thus a wider adoption. Another study surveyed the GAN-based SDG with a particular focus on the utility and privacy of synthetic data [61]. However, the authors did not follow a systematic approach for the selection of studies. Moreover, methods other than GAN were also not considered. Ghosheh et al. [62] reviewed the literature dealing with GANs applications for EHRs with a special focus on data sources and synthetic data. The study also highlighted different evaluation strategies and practical use cases for GAN-generated SD. The authors stressed the need for a standardized mechanism for the utility and privacy evaluation of SD. The study further highlighted the need for transparent reporting and open access availability of relevant resources to facilitate the progress and maturity of the discipline. Some other studies surveyed specific aspects of the SDG such as membership inference attacks against SD [63].

4. Protocol of the Systematic Literature Review (SLR)

Our SLR follows the guidelines of Barbara Kitchenham [64,65] to collect, appraise and synthesize the state of the art in synthetic medical data generation. The objective, research questions, search string strategy, inclusion, and exclusion criteria for conducting this SLR are given in Sections 4.1–4.4.

4.1. Objective

The objective of our SLR is to present the state of the art in synthetic medical data generation for better understanding, consistent representation of important concepts, and aggregation of the growing body of knowledge in this domain.

4.2. Research questions

We investigated the following research questions:

RQ1: *Which different methods have been used to generate synthetic healthcare/medical datasets to address data privacy concerns?*

RQ2: *What design issues/parameters have been considered for privacy-safe synthetic medical data, generation, and what challenges remain to be addressed?*

RQ3: *What primary uses the synthetic medical data have been put to?*

RQ4: *Which methods have been employed to validate the efficacy of privacy-preserving synthetic medical data?*

4.3. Search strategy

We adopted a hybrid search process for this SLR including both automated and manual searches. The automated search was performed on the databases listed in Table 2 on Jul 30, 2022. Following this, we employed a manual forward search of the selected studies. The primary rationale behind forward snowballing is to maximize the coverage and include the most recent works. Hence, to incorporate the latest research under RQ1, we considered forward snowballing a necessary step.

4.3.1. Search string

For the automated search, we formulated a search string constituting the three core elements of our research questions. These include: (1) Synthetic (2) Medical Data (3) Privacy. We explored the synonyms for these and found that a lot of diversity exists in the literature regarding the term "synthetic", which we believe is because this concept is relatively new and lacks standardization [57]. Incorporating the entire spectrum of synonyms for this term resulted in a high percentage of false positives. We settled with the most frequent synonyms which were then cross-checked against our Quasi Gold Standard [64] of 10 known papers in this domain. The finalized search queries are given in Table 2. We also performed double validation for our search string. Firstly, a conventional validation process was employed by ensuring the retrieval of maximal relevant entries in the first 100 results of a google scholar search in the results obtained from the automated searches of the selected databases. We obtained a satisfactory sensitivity score. Secondly, using the Quasi Gold Standard [64], all of the 10 known papers were retrieved by our search string.

4.4. Inclusion and exclusion criteria

4.4.1. Inclusion criteria

The following inclusion criteria were used.

1. Language: Articles written in the English language retrieved against our search string were included.
2. Authenticity: We strictly considered the articles that passed through a peer-review process for inclusion in our review.
3. Validated on Medical data: We included the studies that presented results of the empirical evaluation of their method on medical data.
4. Privacy Consideration: We included only those studies which provided an explicit discussion/evaluation of the privacy guarantees of their methods.

4.4.2. Exclusion criteria

The following exclusion criteria were used.

1. We excluded the articles that did not pass a peer-review process to avoid any misleading analysis.
2. We excluded the studies discussing medical image generation for two reasons; (1) A lot of work has been done for image generation in multiple domains and most of the findings apply to medical images as well, (2) A systematic literature review already exists for medical image generation [66].
3. Our research focuses on medical/healthcare data which possess some unique challenges in modeling and synthetic generation [16], hence we excluded any studies that did not validate their methods on medical data.
4. Researchers argue in [67] that explicit privacy guarantees are required for a synthetic dataset to be privacy risk-free. We excluded the studies lacking the privacy perspective.
5. Studies for which full text was not accessible were also excluded.

4.5. Search results

After applying all inclusion and exclusion criteria, 36 relevant studies were identified. Forward snowballing of the selected studies yielded another 29 papers, and 5 recent studies were included from other sources (suggested by domain experts), mounting the total count to 70 studies included in this review. The details of the screening process are given in Table 2. Fig. 4. shows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram demonstrating the selection process.

Table 2
Search string and results.

Database	Search string	Records retrieved	After duplicate removal	Included after abstract and full text screening
IEEE	(“All Metadata”:“data generat*” OR “synthetic data*” OR “artificial data*” OR “fake data*” OR “synthetic patient”) AND (“All Metadata”:privacy OR confidentiality OR anonymity) AND (“All Metadata”:“medical data*” OR “patient* data*” OR ehr OR emr OR “health* data” OR “care data”)	44	44	5
Pubmed	(“data generation” OR “data generator” OR “synthetic data*” OR “fake data” OR “artificial DATA” OR “synthetic patient” OR “synthetic record”) AND (privacy or anonymity OR confidentiality) AND (ehr OR emr OR “patient data*” OR “medical data*” OR “health data” or “care data” or “hospital data”)	56	56	12
Springer	(“data generation” OR “data generator” OR “synthetic data*” OR “fake data” OR “synthetic patient” OR “synthetic record”) AND (privacy or anonymity OR confidentiality) AND (ehr OR emr OR “patient data*” OR “medical data*” OR “health* data” or “care data”)	1677	1632	15
ACM	[[All: “data generation”] OR [All: “data generator”] OR [All: “synthetic data*”] OR [All: “fake data”] OR [All: “synthetic patient”] OR [All: “synthetic record”]] AND [[All: privacy or anonymity] OR [All: confidentiality]] AND [[All: ehr] OR [All: emr] OR [All: “patient data*”] OR [All: “medical data*”] OR [All: “health* data”] OR [All: or] OR [All: “care data”]]	256	256	4
Forward snowballing and known studies	–	–		34
Total				**70**

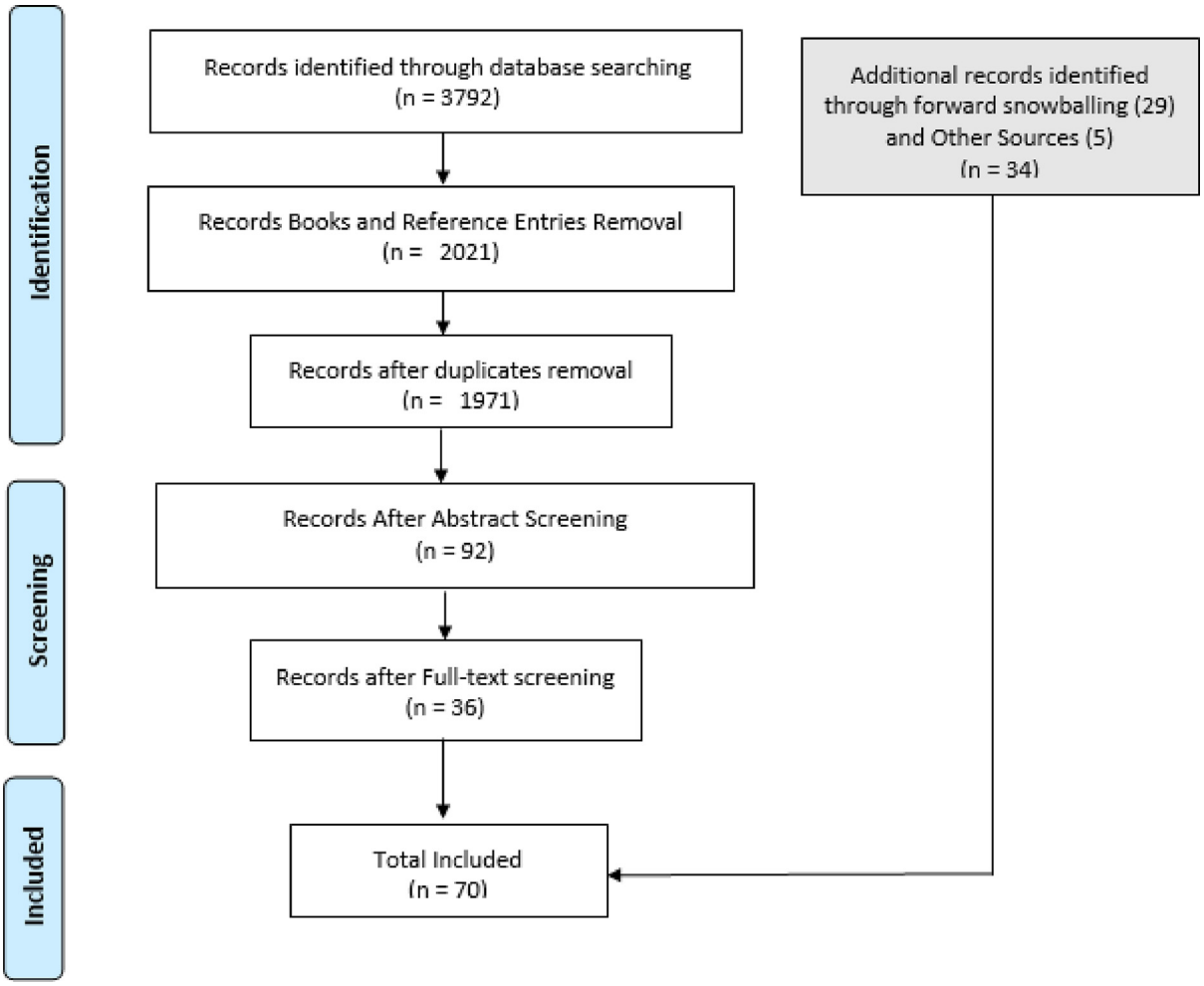

Fig. 4. PRISMA diagram.

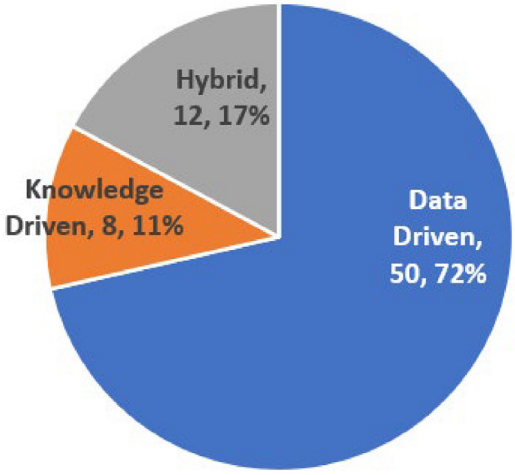

Fig. 5. Classification of studies according to SDG approach.

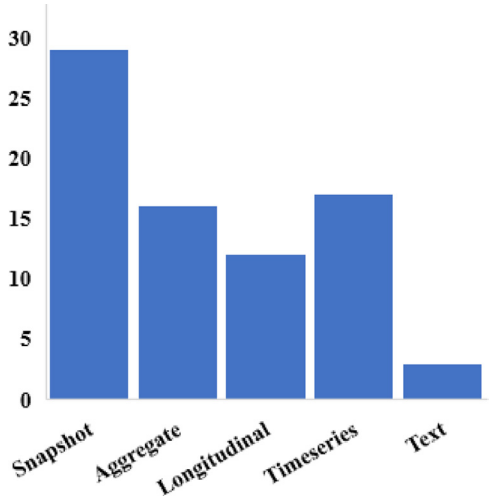

Fig. 6. Synthetic datasets classification according to granularity. Some studies generated multiple datasets of different granularity.

5. Results

Synthetic medical data is required to exhibit the same characteristics as the original data without leaking any sensitive information about the individuals. The quality of synthetic data, however, greatly depends upon the quality of domain knowledge represented in the generative model. A generative model loaded with relevant information is better able to produce good-quality synthetic data with a high incidence of ground truth. In the following subsections, we aim to provide answers to our research questions. Section 5.1 discusses the core concerns and challenges in synthetic medical data generation, in response to RQ2, Section 5.2 presents a narrative review of the included studies, and Section 5.3 highlights the quality evaluation metrics proposed in the literature for synthetic medical data.

5.1. Major challenges and concerns in synthetic medical data generation

Below we highlight the major concerns and challenges in synthetic medical data generation, in response to our RQ2 and as an underpinning to orchestrate the narrative review in subsequent sections. We have identified seven major aspects that need to be considered for synthetic medical data generation. Sections 5.1.1–5.1.7 discuss these and the associated challenges.

5.1.1. Ground truth sources (KD and hybrid approaches)

The KD approaches utilize public bodies of knowledge to extract ground truth to construct the generative model. Though abundant and freely available, the knowledge exists in heterogeneous formats and demands extensive human effort to extract, process, and consolidate meaningful information amenable to formalizing the generative model. The selection of appropriate ground truth is a crucial factor in determining the overall quality of generated SD.

5.1.2. Dataset granularity (DD and hybrid approaches)

The DD generation requires quality (real) datasets to train the generative model. The EHRs collected at health care sites are usually multi-dimensional and longitudinal datasets with fine-grained features, recording patient history over several hospital encounters. This data provides a wealth of information for day-to-day medical practice but is restricted for secondary use under privacy laws. Although many anonymized medical datasets are available for the researchers (Table B.2 - Appendix B), most of them offer only a coarser representation of features. The quality of SD produced by a DD model is capped by the quality of available training data. Based on the granularity of patient records we define the following granularity formats for training and synthetic datasets and highlight the challenges associated with the availability and generation of each type.

- *Snapshot or cross-sectional:* This is the most common format for public medical data (Table B.2 - Appendix B) which presents one snapshot of a patient's EHR containing only the selected attributes. Many of the published medical datasets exist in this format. However, they provide a partial view of the overall EHR, and thus the SD generated from such data is generally useful for specific applications. Examples of this format include UCI Heart Disease Dataset [112] and SEER Data set [113].
- *Aggregate:* The longitudinal data of a patient containing multiple encounters are flattened into a single row with each column representing an event (e.g., disease and lab codes). Binary or count aggregations have been utilized by researchers in many studies, which represent the presence or absence (binary) or the number of occurrences (count) of an event in the patient history. Although only a few published datasets (e.g. [114]) originally existed in this format (Table B.3 - Appendix B), many studies aggregated the longitudinal data as a preprocessing step before generative model training and produced a snapshot or aggregated SD. Aggregated datasets provide wider coverage of patient health conditions compared to snapshot data but lose the temporal and sequential relationships between features.
- *Timeseries:* Data from various monitoring devices come in this format. Timeseries data include features recorded repeatedly after certain time intervals e.g., EEG and ECG. Realistic synthesis of timeseries data is an active area of research and poses a considerable challenge [115]. Examples of medical timeseries datasets include UCI EEG Eye State dataset [112]. A few of the included studies targeted this type of SD as shown in Fig. 5.
- *Longitudinal:* True longitudinal data consists of multiple attributes recorded at different points in a patient's lifetime. Longitudinal data is the richest and most detailed representation of patient EHRs. However, because of the sensitivity of the information contained in them, these datasets are rarely available for public use. MIMIC [116] is an example of a

Table 3
Knowledge-driven SDG methods (Feature Types M: Mixed, C: Continuous, D: Discrete | Synthetic Data Granularity: S: Snapshot, A: Aggregate L: Longitudinal).

Study	GT source	GM	SD set granularity	Feature types	Realism validation	Strengths	Limitations
[68]	Biomedical, Publications, Web resources, Medical experts	PPDL document	S	M	Utility evaluation for clinical trial selection.	Simplistic formulation of GM. Interoperability with clinical support systems.	Lacks comorbidity representation. Configuring rules for longitudinal data is challenging. Defining rules for diversity in SD is hard
[36]	Biomedical publication, Web resources, Domain experts	PPDL documents	S	M	Utility evaluation as test data for a clinical support system Medical expert inspection	Simplistic formulation of GM. Interoperability of the synthetic data with other systems.	Fine-grained data synthesis is intricate and error prone.
[69,70]	Local health department publications, Medical research and academic literature, Web resources, Domain experts	State Transition Machines (Synthea)	L	M	Expert inspection	Fully qualified Longitudinal EHR. Extendible for many phenotypes and scenarios.	Dependence upon the relevant CPGs Generated Data is not suitable for clinical decision making. Multimorbidity is difficult to model
[71]	Local health department publications, Medical research and academic literature, Web resources, Domain experts	State Transition Machine (Synthea)	L	M	Comparison against reference statistics (Incidence and Prevalence)	Fully specified EHR generation. Extendible for many phenotypes.	Dependence upon the relevant CPGs. Generated Data is not suitable for clinical decision making Multimorbidity is difficult to model.
[25]	Local health department publications, Medical research and academic literature, Domain experts	State Transition Machines (Synthea)	A	M	Comparison against reference statistics (Incidence and Prevalence)	Same as [71]	Same as [71]
[72]	Health incidence statistics synthetic mass (KD synthetic data from Synthea)	State Transition Machines (Synthea)	L	M	Comparison against reference statistics (quality of care)	Longitudinal Patient Records Flexible and extendible for other phenotypes	Same as [71]
[32]	Medical research publications Domain experts Health Incidence Statistics and Prevalence	State Transition Machines (Synthea)	L	M	Comparison against reference statics	Longitudinal Patient Records Flexible Modeling of Contagious Disease	Same as [71]

de-identified longitudinal dataset. Synthesis of longitudinal SD through DD approach requires access to longitudinal real data which is hard to approach because of privacy considerations. Thus, we see fewer DD based methods for longitudinal SD, but a majority of the KD based methods support this format (Table B.3 – Appendix B). Fully furnished synthetic longitudinal data generation through DD methods is still an open research problem.

- *Text/Clinical Notes:* Clinical notes are an integral part of EHR and provide important information for day-to-day clinical practice. We classify this type as an independent category in the context of SD because of the specialized methods for natural language processing, which are significantly different from those for relational data.

The provision of an appropriate level of detail in the training data is an important factor affecting the utility of generated SD. Some applications might work well with only the aggregated information, for example, estimating resource requirements for a hospital. Others such as medical decision support systems will require detailed data. Tables 3, 4, and 5 highlight the granularity formats of synthetic data generated by included studies.

5.1.3. Feature types

Medical data consists of continuous as well as discrete features which support different kinds of analyses and applications. Representation of a feature in the appropriate type is essential for better realism and higher utility of SD for the target task. However, we noted that most of the earlier DD generative models dealt with only one type of feature i.e., either discrete or continuous, which limits the applicability of this data to specific use cases. Numerous studies employed data transformation techniques to convert between feature types to allow for mixed features in the final dataset. However, the conversions are usually lossy and add additional noise to the generated SD. A few DD generative models provide built-in support for mixed feature generation. For KD generative models, feature type is a matter of choice for the model curator which is manifest from the large proportion of KD approaches incorporating mixed feature generation (Table B.4 - Appendix B – Highlighted in Bold). Types of features supported by an SDG is an important consideration in selecting an appropriate method for the target application.

5.1.4. Generative model

A generative model forms the core of any synthetic data generation method and significantly influences the quality of the generated data. In addition to the intrinsic challenges of medical SD generation (e.g., high dimensionality), different generative models may face specific technical difficulties, such as mode collapse with GANs [117], which can affect the quality of synthetic data. We consider the selection of an appropriate generative

Table 4
Data driven synthetic data methods employing classical Generative Models (Feature Types — C: Continuous, D: Discrete, M: Mixed) | (Synthetic Data Granularity — A: Aggregate, L: Longitudinal, S: Snapshot).

Study	GT/Training data	GM	SD granularity	Feature types	Realism metrics	Privacy metrics	Strengths	Limitations
[73]	UCI Machine Learning Repository (Thyroid and Diabetes)	Mixture of Multivariate Normal (MVNs)	S	C	Moments Based Comparison Distinguishability Metric	v-dispersion criteria	Flexible GM to capture any complex multivariate distribution.	Lacks support for discrete values. Lacks support for longitudinal data. Minimal scalability for larger volumes of high-dimensional data.
[74]	UCI Machine Learning Repository (Thyroid), California Neoplasm Dataset	Mixture of Multivariate Normals (MVNs)	S	M	Moments Based Comparison Distinguishability Metric Parameter Agreement	Probabilistic k-anonymity	Flexible GM to capture complex multivariate distributions. Support for Mixed data synthesis.	Lacks support for longitudinal data. Minimal Scalability for high-dimensional data
[75]	National Long Term Care Survey (NLTC)	Bayesian Networks	S	D	Dimension-Wise Prediction Total Variational Distance	Differential Privacy	Strong privacy guarantees Suited to high-dimensional medical data.	Lacks support for temporal and longitudinal data.
[42]	Indian liver patient dataset (ILPD)	T-Copula	S	M	Visual Plausibility (Feature correlation heatmap, Pair-Wise feature distributions) Uni and Multivariate distance metrics Performance Agreement, NMDS similarity	Reproduction Rate – Record Duplication	Simplistic GM Modeling Copula-based GM better captures multivariate dependencies. Supports mixed feature types.	Lacks support for temporal and longitudinal data.
[76]	BSA Inpatient Claims PUF	MCMC Gibbs Sampling	S	D	Parameter Agreement Visual Plausibility (Distribution Histogram)	Synthetic l-diversity $\in$ differential perturbation	No assumption regarding the probability distribution of real data. Flexible to adapt to complex density functions.	Discrete features only Produces partially synthetic data. A tradeoff between data truthfulness and privacy
[77]	Patient Discharge Data California	MCMC Gibbs Sampling	S	D	Rank Correlation Coefficient Parameter Agreement Univariate Distributional Distance	Synthetic l-diversity $\in$ differential perturbation	No assumption about the probability distribution of real data. Can work with complex distributions	Discrete features only. Produces partially synthetic data.
[23]	St Louis Pediatric Intensive Care Data	Kernel Density Estimation	A	M	Dimension Wise Statistics, Comparison against reference statistics	Outlier Similarity	Mixed feature types. Model simplicity	Does not support longitudinal and timeseries data
[78]	COVID N3C	Kernel Density Estimation	S	M	Comparison against reference statistics Performance Agreement (Classification)	Default	Same as above	Same as above
[79]	COVID N3C	Kernel Density Estimation	S	M	Dimension Wise Statistics Comparison against reference Statistics	Default	Same as above	Same as above
[80]	MIMIC UCI – Heart Disease, Diabetes	Bayesian Networks	A S	D M M	Dimension Wise Statistics and Prediction, Association Rule Mining, Visual Plausibility (Heatmap, Distribution Histograms), Distinguishability	Nearest Neighbor Similarity Metric	Model Interpretability. Computationally efficiency. Mixed Features Support	Suboptimal performance with discrete data. Unable to capture the temporal characteristics.
[42]	CPRD Aurum database	Bayesian Networks	S	M	Expert Inspection, Augmentation Test Visual Plausibility – (Correlational Matrix, Distribution Histogram)	Outlier Similarity	Model Interpretability. Computational Efficiency. Mixed Feature Support.	A general increase in the inter-feature correlation strength as compared to real data.

Table 5
DD Methods employing ML based Generative Models (Feature Type: M: Mixed, C: Continuous, D: Discrete | Synthetic Dataset Format- L:Longitudinal, S: Snapshot, A: Aggregate, T: Temporal) [the M* means that the method generates mixed features through data transformation but the generative model itself might not support mixed feature generation].

Study	GT/Training data set	GM/Base model	SD granularity	Feature types	Realism validation metrics	Privacy metrics	Strengths	Limitations
[81]	MIMIC-III	MedGAN/ GAN	A	D	Dimension Wise statistics. Dimension Wise Prediction. Expert Inspection	KNN attribute estimation metric	First to work with discrete health data generation.	Lacks support for mixed data types. Aggregated representation of patient diagnoses only. Favors univariate realism.
[82]	MIMIC-III	MedGAN/ GAN	A	D	Visual Plausibility -Distribution Histograms	Inherited	Extended medGAN to accommodate demographics	Lack of support for mixed data types.
[83]	CDC - NCHS Health Dataset	TableGAN/ DCGAN	S	M	Response Agreement Visual Plausibility – Distribution Histograms	Nearest Neighbor Metric	Supports mixed data type synthesis. Pure tabular data	Lacks support for temporal and longitudinal data
[84]	UCI Dataset, MIMIC	DP-CTGAN/ CTGAN	S	M	Performance Agreement – Classification	Differential Privacy	Supports mixed data synthesis. Tabular format	Realism is low for practical applications
[85]	MIMIC-III	RSDGM/ WGAN	A	M	Visual Plausibility – Distribution Histogram, Correlation Heatmap, Feature wise distance.	Inherited	Support for mixed feature types Lab codes included	Validated on small/low dimensional dataset.
[86]	MIMIC-III NHIRD, Taiwan	MedBGAN/ BGAN	A	D	Dimension Wise Statistics Dimension Wise Prediction KS-Similarity Test Association Rule Mining	Inherited	Improved model stability and performance against benchmark	Model not validated on mixed feature type generation. Lacks support for longitudinal data.
[87]	MIMIC-III Extended MIMIC-III NHIRD, Taiwan	MedBGAN/ BGAN medWGAN/ WGAN-GP	A	D	Dimension Wise Statistics Dimension Wise Probability KS- Similarity Association Rule Mining	Inherited	Improved model stability and performance against benchmark	Model not validated on mixed feature type generation. Lacks support for longitudinal data.
[22]	SEER's dataset (Breast Cancer, Respiratory Cancer, Leukemia)	MC-MedGAN/ medGAN	S	D	KL Divergence Dimension Wise Probability Visual Plausibility-Correlation heatmap, Distribution histogram Log Cluster Support Coverage Domain Rule Preservation	Identifiability Nearest Neighbor Attribute Estimation	Better privacy against benchmark models.	Inability to capture the multivariate structure of real data.
[43]	Private Pediatric Data	GcGAN/ T-wGAN	A	D	Dimension Wise Statistics Feature Co-occurrence frequency Performance Agreement – Classification	Inherited	Enriched records with medication information.	Lacks support for mixed feature types. Temporal dependencies lost.
[39]	MIMIC-III	SC-GAN/ GAN	T	M	Dimension Wise Statistics Visual Plausibility - Correlation Heat Map Data Augmentation Tests Expert Inspection	Inherited	Generates timeseries data for multiple patient visits. Includes medication dosage information.	
[88]	MIMIC-III UCI Epileptic Seizure Recognition	Cor-GAN/CNN	A T	D C *	Dimension Wise Statistics Dimension Wise Prediction Performance Agreement	Nearest Neighbor Similarity Metric	Greater similarity in correlational structure of real and synthetic data. Learns temporal characteristics for continuous data.	Method not demonstrated on mixed feature type generation. Lacks support for longitudinal data
[89]	MIT-BIH Arrhythmia Database	SynSigGAN/ GAN	T	C	Distribution distance-based metrics Visual Plausibility – Distribution Histogram Pairwise correlation coefficient	Inherited	Learns the temporal structure of signal data with high realism.	Lacks support for longitudinal EHR.

(*continued on next page*)

Table 5 (*continued*).

Study	GT/Training data set	GM/Base model	SD granularity	Feature types	Realism validation metrics	Privacy metrics	Strengths	Limitations
[90]	VUMC Synthetic Derivative	EMR-WGAN EMR-CWGAN	A	D	Dimension Wise Statistics Dimension Wise Prediction First Order Proximity Metric Latent Space Metrics	Nearest Neighbor Similarity Metric Reproduction Rate	Improved statistical realism for aggregate synthetic data	Lacks support for continuous features. Temporal features are lost.
[91]	VUMC Synthetic Derivative	HGAN/ EMR-CWGAN	A	M	Dimension Wise Statistics Dimension Wise Prediction Cross-Type Conditional Distribution Statistics Association Rule Mining Domain Rule Preservation	Nearest Neighbor Similarity Metric KNN Attribute Estimation	Supports mixed feature types.	Lacks support for longitudinal and timeseries data synthesis.
[40]	Cerner Health Facts database	SMOOTH-GAN/ WGAN-GP, CGAN	A	M	ML Performance Agreement DM Congruence – feature Ranking	Maximum Mean Discrepancy	Supports Mixed feature generation. Can generate records with specific characteristic.	Temporal features are lost.
[41]	Gene Expression Data	GEG/ WGAN	S	C	Data Augmentation Test	Inherited	Gene expression data generation.	Limited scalability of model for large synthetic dataset.
[92]	Private dataset of medical text	mtGAN/ RNNs,CNN	Text	D	Data Augmentation Test Distinguishability Metric	Default	Medical text generation	Susceptible to attribute disclosure.
[28,38]	MIMIC-III	Health-GAN/ WGAN-GP	A S	M*	Nearest Neighbor Adversarial Accuracy Visual Plausibility - PCA Utility for education and research	Nearest Neighbor – Privacy Loss	Enables exporting the model for on-the-fly data synthesis.	Strict privacy evaluation required before exporting the model.
[93]	Kaggle Medical Cost Dataset, PIMA Indian Diabetes Dataset	pGAN/ GAN	S	C	Nearest Neighbor Adversarial Accuracy Metric Performance Agreement	Nearest Neighbor Adversarial Accuracy – Privacy Loss	–	Limited scalability for high dimensional datasets and mixed features.
[80]	MIMIC-III	Health-GAN/ WGAN-GP	A	M*	Nearest Neighbor Adversarial Accuracy Performance Agreement	Nearest Neighbor Adversarial Accuracy– Privacy Loss Membership Inference Risk	–	–
[94]	MIMIC-III	Health-GAN/ WGAN-GP	T	M	Performance Agreement	Inherited	Learns temporal characteristics of data	Substantial preprocessing of training data is needed.
[95,96]	Autism Spectral Data	HealthGAN	T	C	Root Mean Squared Error Pearson Correlation Coefficient Directional Symmetry Short Timeseries Distance	Inherited	–	–
[97]	UNOS-Heart waitlist Kaggle cervical cancer dataset UCI Epileptic Seizure Recognition	PATE-GAN/ GAN, PATE Frame-work	S	M	Performance Agreement Synthetic Rank Agreement Feature Ranking Agreement	Differential Privacy	Supports mixed feature generation. Built-in differential privacy.	Does not support longitudinal and temporal data generation.
[98]	MIMIC-III	PPGAN/ WGAN	A	D	Generate score	Differential Privacy Moment Accountant Metric	Strong privacy preservation mechanism	Loses temporal features. Lacks longitudinal data support.

(*continued on next page*)

Table 5 (*continued*).

Study	GT/Training data set	GM/Base model	SD granularity	Feature types	Realism validation metrics	Privacy metrics	Strengths	Limitations
[99]	MIMIC-III SPRINT Dataset	AC-GAN/ GAN	S T	C	Moments Based Similarity Pairwise Correlation Performance Agreement Expert Inspection Feature ranking agreement Parameter congruence	Differential Privacy	Learns temporal trends in the data. Strong privacy guarantees	Lacks support for mixed feature types. Poor scalability for high-dimensional data.
[100]	Philips eICU Database IKD EEG Database	PART-GAN/ AC-GAN, CGAN	T	C	Inception Score Visual Plausibility – Distribution Histograms	Differential Privacy Nearest Neighbor Metric	Improved stability and convergence during training. Better utility of SD under DP	Limited scalability for high-dimensional datasets.
[101]	MAGGIC Dataset UNOS Dataset	ADS-GAN/ WGAN-GP, CGAN	S	M	Dimension Wise Statistic Visual Plausibility – Covariance multivariate Distributional Distance Performance Agreement	Identifiability Loss	Stronger multivariate relationship preservation in SD than benchmark models. Strong privacy guarantees.	Lacks support for longitudinal data.
[102]	Private Data (Fitbit)	BGAN	A	M	Dimension Wise Statistics Visual Plausibility – Distribution Histogram	Differential Privacy	Provides control over privacy settings. Supports mixed features generation.	Limited scalability for high-dimensional datasets.
[103]	MIMIC-III	DAAE/ GAN, AE	L	D	Expert Inspection	Differential Privacy	Generates longitudinal records.	Lacks support for mixed feature types. Medication and procedure codes not considered
[104]	MIMIC III UCI datasets	RDP-CGAN/ CNN, CAE	A S	M	Maximum Mean Discrepancy Dimension Wise Prediction	Renyi Differential Privacy	Captures correlations well. Provides tighter privacy bounds than DP.	Longitudinal data synthesis not supported.
[105]	VUMC-SD	SynTEG/ EMR-CWGAN	L	D	First Order Temporal Statistics Performance Agreement Latent Space Metric	Perplexity Distributions Attribute Estimation Likelihood	Supports longitudinal data generation.	Mixed feature types not supported.
[106]	VUMC-SD, NIH All of Us- Dataset	LS-EHR/ GAN, RNN	L	D	Distinguishability Metric	Inherited	Supports longitudinal data generation.	Mixed feature types not supported.
[107]	MIT-BIH Arrhythmia Database BIDMC PPG and respiration database EEG eye state database Balleostrocardiogram database	BirNN/ RNN	T	C	Pair-wise Pearson Correlation Coefficient Distributional Distance	Inherited	Can generate signal data. Learns temporal trends of the real data.	Lacks support for mixed feature types.
[108]	UCI Breast Cancer UCI Diabetes UCI Mammographic Mass	DP-SYN/ AE	S	M	Performance Agreement – Classification Total Variational Distance	Moments Accountant Metric	Performance not deteriorated with imbalanced datasets.	Suitable for low dimensional datasets. Lacks support for longitudinal records.
[109]	Private Longitudinal dataset	EVA/AE, conditional AE	L	D	Performance Agreement – Classification Augmentation Test Expert Inspection	Reproduction Rate	Generates longitudinal data.	The patient records have fixed number of visits. Timestamping of the visits not supported. Assumes fixed gaps between visits.

(*continued on next page*)

Table 5 (*continued*).

Study	GT/Training data set	GM/Base model	SD granularity	Feature types	Realism validation metrics	Privacy metrics	Strengths	Limitations
[31]	UCI Breast Cancer Dataset UCI Parkinson's Dataset UCI Diabetes Dataset	Random-ized Decision Trees	S	M	Performance Agreement – Classification, Regression Feature Ranking Agreement	Default	Efficient generation of large volumes of synthetic samples.	High space complexity of the GM. Mixed feature types not supported.
[110]	Project Data Sphere Datasets	Decision Trees	S	M	Univariate Distribution Distance Dimension Wise Prediction Performance Agreement Distinguishability Metric Feature Ranking Agreement	Default	Efficient generation of large volumes of synthetic samples (low dimensional). Robust to missingness in training data.	Suitable for low dimensional data sets.
[111]	Canadian COVID-19 case dataset	Decision Trees	S	M	Univariate Distribution Distance Dimension Wise Prediction Performance Agreement Distinguishability Metric Feature Ranking Agreement Comparison against Reference statistics	Standardize Mean Discrepancy (SMD)	Same as [110]	Same as [110]

model a crucial factor regulating the applicability of the generated SD for a given scenario.

5.1.5. Realism validation

Realism is the primary quality attribute of synthetic data. Researchers have proposed a range of different realism metrics. It is important to evaluate a synthetic dataset for its statistical validity as well as its utility as a substitute for the real data in the target application. Suitable realism metrics must be employed to thoroughly evaluate the SD before releasing it for use. Realism is among the most important factors influencing the selection of a particular method for the given scenario.

5.1.6. Privacy preservation

Privacy is a concern for the DD and HBD approaches that utilize private data for model training. Fully synthetic approaches do not face the identity disclosure risk as no direct mapping to a real record is present. However, an adversary, equipped with relevant background knowledge, may gain membership or attribute information about real data from the published synthetic data. It is important to ensure the resilience of SD to disclosure risk before publishing.

5.1.7. Computational efficiency

This aspect has not been focused much on in the literature primarily because SD generation is primarily presented as an offline procedure. This concern will become relevant as the field matures and SD finds wider adoption.

5.2. Synthetic medical data generation: Methods and models

In this section, we present a narrative review of the included studies with a focus on the SDG concerns highlighted above to answer our RQ1. We structure the discussion based on the SDG approach employed by each study. It can be seen in Fig. 5 that the frequency of DD studies is significantly higher compared to others. Section 5.2.1 discusses the KD methods followed by a discussion of data-driven and hybrid techniques in Sections 5.2.2, and 5.2.3 respectively.

5.2.1. Knowledge driven synthetic medical data generation

Knowledge-driven techniques make use of publicly available knowledge and thus pose minimal privacy risks. However, correct, complete, and consistent representation of knowledge is arduous and requires considerable manual effort. Proponents of knowledge-driven techniques argue against data-driven methods for not only posing higher risks of privacy attacks but also facing the bootstrap or the "chicken or egg" paradox of data availability [118]. Moreover, real-world biases might also seep into the generated data [81]. Thus, we see fewer studies in this category (Fig. 5, Table B.1, Appendix B). Depending upon the nature of the generative models employed, we classify the KD approaches as (1) Structural and (2) Behavioral methods. The primary difference between the two approaches is that the former adopts a structural modeling approach in form, or rules defining constraints on the structure of the data, while the latter makes use of dynamic models to formalize the behavior of entities that produce data. A summarized view of the KD methods with respect to the sources of ground truth, generative model employed, feature types supported, and realism evaluation metrics are presented in Table 3. We also highlight the major strengths and limitations of the methods.

A. Structural Methods

These methods focus on the structural aspect of the data to be generated and employ static models for GMs. The domain knowledge is formalized as rules or constraints that define the structure of the data. Advanced Patient Data Generator (APDG) [68] embodies one of the simplest SDG techniques employing an XML-based Patient Data Definition Language (PDDL) to define data generation rules. These rules impose constraints on the structure, values, and correlational dependencies of the synthetic dataset. The utility of generated data is tested in a clinical trial selection scenario with promising results. The simplicity of this technique makes it a convenient choice for practitioners, however, in its present scope, it lacks support for comorbidity which is an important factor in clinical practice. Moreover, configuring rules for longitudinal records is labor-intensive and susceptible to incompleteness, inconsistencies, and lack of diversity. Another study [36] tested the efficacy of APDG for generating synthetic data for patients with a first diagnosis of depression, to be used as test data for a clinical decision support system. The authors

claimed satisfactory performance of SD for the selected task but failed to present actual results.

B. Behavioral Methods

These methods employ behavioral models such as state charts or computer programs to formalize the domain knowledge for the generative model. The models in this category formalize the dynamics of entities instead of individual data units as in the case of structural models. A prominent work in this category is Synthea [71] which the authors call a "cradle to grave" EHR generation. Several research studies [32,69–72] presented the method at different stages of its development. Based on the ATEN framework [69] for SDG, Synthea supports longitudinal patient record generation for the selected phenotype. The core architecture is based on the Publicly Available Data Approach for the safe generation of the Realistic Synthetic (RS) EHR (PADARSAR) approach [119] that derives the requisite ground truth from publicly available knowledge only, thus eliminating disclosure risk. Content modeling is carried out by a component CoMSER (Content Modeling for Synthetic E-Health Records) [70] which extracts constraints from Clinical Procedural Guidelines (CPGs), Disease Progression Models (DPMs), and Health Incidence Statistics (HIS). Clinical care pathways called Care Maps (CMs) are developed from these constraints and other sources of clinical knowledge and formalized as State Transition Machines (STMs) to enable the generation of synthetic records. CMs include various types of states to represent the different health events taking place during hospital encounters (such as the onset of a medical condition, prescriptions, labs, and death) and the corresponding progression of the health condition guided by the transition constraints. CMs aim to model the entire medical life span of a patient from the first hospital encounter to the last marked by death and thus called "cradle to grave" modeling. Traversals through the care map STMs yield synthetic records. The method was employed for the generation of the Type 2 diabetes synthetic dataset in [71]. However, significant discrepancies were observed compared to the real incidence and prevalence statistics, particularly for the comorbid complications. Liu et al. [25] also generated Type 2 diabetes data using Synthea modules to support urban planning for improved living. Again, some discrepancies were observed between the real and synthetic prevalence statistics. Nevertheless, the generated data carries satisfactory utility for the intended objective i.e., support for urban planning. A synthetic dataset for child delivery episodes was generated in [70] and a team of domain experts concluded that the synthetic timelines and clinical notes manifested sufficient realism. Later, a study [72] employed quality of care as a metric to validate realism in Synthea-generated synthetic data. Results revealed significant discrepancies in the post-intervention outcomes of the real and synthetic data for certain phenotypes. The synthetic data generally showed a higher incidence of certain conditions in response to some interventions. Nevertheless, being KD, the anomalies in the generated data can safely be attributed to omissions and errors in the generative modeling which can be addressed by furnishing the missing details incrementally. Good for various analyses, the data from Synthea, is generally not suitable for the critical clinical decision making. Another study [32] utilized Synthea to generate synthetic COVID-19 patient records to support the development and validation of various solutions during the corona pandemic. This data has also been used for hackathons and conferences [32]. Synthea carries the promises of KD-based generation by leveraging complete control over the generation process which allows for further extensions and adaptations to meet specific needs. However, the quality of generated data greatly depends upon the quality of the generative model i.e., the comprehensiveness, correctness, and completeness of the care maps and corresponding simulation scripts.

5.2.2. Data-driven synthetic medical data generation

These techniques rely on a generative model laden with the necessary knowledge extracted from real data to generate new samples. This type of SDG has been most widely explored by researchers, manifested by the high proportion of studies in this category (Fig. 5, Table B.1 – Appendix B). We classify the DD approaches as (1) Classical Methods and (2) Machine Learning Based Methods which we describe below. Tables 4 and 5 present a summary of DD methods.

A. Classical Methods

One of the oldest approaches to generating artificial data is the classical statistical method. Synthetic data has been a popular Statistical Disclosure Control (SDC) method since 1993 [48]. However, SDC for the most part deals with partially synthetic data. The fundamental principle underlying the classical approaches is to estimate the distribution of the original data and derive new synthetic samples from it. However, most of these methods are based on parametric estimation which makes some assumptions about the shape of the data distribution. This limits the potential of the generative models. Moreover, the high-dimensionality and complex correlational structure of medical data pose considerable challenges in the estimations. Simulation models are expressed as inferred, conditional, joint, or marginal distributions learned from the sample of real data. We note that most of the studies in this category dealt with snapshot or aggregate datasets generation only. Synthetic data produced in this manner is useful for certain verification and planning tasks but is not appropriate for clinical analysis [120]. Several classical models have been used by researchers in this regard. These are (1) Mixture Models, (2) Copulas, (3) Monte Carlo, and Bayesian Network.

1. *Mixture Models:* Simply stated, a mixture model is a collection of simpler data distributions to estimate a complex density function where each mixture component represents the density of one cluster or subpopulation. Medical data inherently consists of patient subpopulations and mixture models are a natural fit for modeling such data. The mixture of Gaussians or Multivariate Normal (MVNs) used in [73] and later in [74] by Oganian et al. provides a flexible approach to modeling a medical dataset by incorporating into the mixture, the estimated MVNs for each cluster. Samples are then derived from this MVN mixture model under a privacy rule termed v-dispersion criteria [73], which regulates the spread of the synthetic clusters to control disclosure. Comparisons against other similar approaches yielded a better performance for various statistical realism metrics. However, actual results from the experiment were missing in the first study [73]. Moreover, this study failed to demonstrate the method for discrete variables. The discrepancies were addressed in a later study [74] by the same authors with results from more rigorous experimentation for continuous, discrete, and mixed data types generation validated against an array of realism metrics (Table 4)
2. *Copulas:* Copulas are another mathematical tool that couple a multivariate distribution to its univariate marginals. Copulas have found extensive application in simulating the linear or nonlinear relationships among multivariate data in scientific and engineering studies and are considered appropriate for the modeling of complex multivariate medical data as well [121]. In a recent study, Wang et al. [42] demonstrated the results of the evaluation of synthetic data generated using the t-copula. The authors concluded that copula-based synthetic data can preserve the uni and multivariate correlation structure of the original data with acceptable privacy guarantees. The study also

demonstrated harmonious clustering trends between real and synthetic data on the Non-metric Multi-Dimensional Scaling (NMDS) plots, which further confirms good realism. However, the study employed a superficial privacy assessment by ruling out any synthetic rows that duplicated any real records. A more rigorous privacy assessment is desirable.

3. *Statistical Simulation/Monte Carlo (MC)* is among the most popular statistical simulation approach for non-trivial distributions. Gibbs Sampler is a Markov Chain Monte Carlo (MCMC) method that derives a sequence of observations approximated from a specified multivariate probability distribution in situations when direct sampling is prohibitive. Yubin et al. employed Gibbs approach through a sequence of conditional distributions to approximate the joint distribution of the high dimensional medical data in [76,77]. The method produces partially synthetic data like multiple imputations. Feature values are generated sequentially through conditional distributions given the synthetic feature from the earlier iterations. The authors apply hashing on the feature vectors for space and computational efficiency. Privacy guarantees under the differential and l-diversity criteria are provided through perturbation of the estimated conditional distributions, which adds an appropriate level of noise in the generative model. The resulting synthetic data resembles an anonymized dataset produced through generalization and suppression operations. The authors validated the utility of data through predictive modeling [76] and correlational metrics [77]. However, the method presently supports discrete features and partially synthetic data generation only.
4. *Bayesian Network (BN)* is a well-suited choice for medical data for its interpretability and ability to learn from small datasets. PrivBayes [75] is a prominent work for differential privacy generation of synthetic data using a Bayesian Network. A particular highlight of the method is the support for high-dimensional data which is an inherent characteristic of medical datasets. The method constructs a BN and extracts low dimensional marginals from it which are then injected with noise for Differential Privacy (DP) and finally SD is derived from this set of noisy marginals. While this approach provides a significant gain in computational tractability, the quality of synthetic data greatly depends upon the accuracy of the low dimensional approximations of the marginals. Another recent study [80] employed a non-parametric approach using an Acyclic Bayesian Network (ABN) extracted from real data. A top-down path traversed in this BN yields one complete synthetic record. The authors employed several uni- and multivariate metrics to validate the realism with promising results. However, discrepancies in the temporal dependencies of the features were observed in the SD. Moreover, the sub-optimal performance of the model in capturing continuous data was noted. The overall realism scores showed a general improvement in performance over a benchmark model (medBGAN) [86]. The authors highlighted the ability of BNs to capture rare features in original data adequately well. Another research [42] studied the plausibility of BNs for synthetic Cardio Vascular Disease (CVD) data. Random sampling through the extracted BN is employed to generate a synthetic sample. The evaluation of synthetic records showed good uni- and multivariate similarity. However, the inter-feature correlation strengths generally appeared stronger in synthetic data than in real values. Qualitative assessment through NMDS plots revealed unison in clustering trends between real and synthetic samples. Furthermore, two medical experts confirmed sufficient indistinguishability between synthetic and real records generated by this method.
5. *Kernel Density Estimation (KDE)* is a non-parametric method to estimate the distribution of a population from a finite-sized sample. MDClone [122] is a data analytics platform for health data that supports synthetic data generation using a modified multivariate KDE method. Being a non-parametric approach, it is flexible and can be applied to complex distributions, which makes it suitable for health data. Foraker et al. [23] evaluated the utility of MDClone-generated SD for clinical (sepsis prediction) and non-clinical (data analysis) applications and concluded that SD can accelerate research and development by substituting hard-to-get real data. Other uses of MDClone SD appear in [23,79] for the United States National COVID Cohort Collaborative (N3C) dataset. The studies showed a high resemblance of MDClone SD with the ground truth in terms of summary statistics and geo-spatial data analysis outcomes [78,79]. Study [78] also carried out a utility assessment of SD for 14-day admission risk prediction for COVID with results comparable to the real data. However, a formal evaluation of the privacy of the generated data was lacking [23,78,79].

B. Machine Learning Based Synthetic Medical Data Generation

Machine learning has recently gained a lot of attention for data-driven synthetic medical data generation. We group the techniques by the ML models employed for generative modeling into four categories: (1) Generative Adversarial Networks (GANs), (2) Neural Networks, (3) Autoencoders and (4) Decision Trees. The following subsections present our narration of the studies employing ML-based generative models. A summarized view of these studies including the dataset, generative model, synthetic data granularity, and realism and privacy metrics is presented in Table 5.

1. Generative Adversarial Networks:

Generative Adversarial Networks (GANs), proposed by Ian Goodfellow in 2014 [123], are considered a breakthrough in AI owing to their remarkable performance in various supervised and unsupervised learning applications. Since their inception, GANs have gained a lot of importance in medical informatics [124]. GANs are deep learning models that provide implicit modeling of the complex multidimensional distribution of training data which can be used to derive new synthetic data points having the same distribution. GANs consist of two components: a generator that attempts to generate realistic but fake data and a discriminator that aims to distinguish between the generated fake and the real data. By competing against each other, the generator tends to learn the distribution of the real data to fool the discriminator to take it for real data while the discriminator learns to distinguish meticulously between real and fake samples. GANs have been shown to outperform competitors like Autoencoders (AE) and Pixel RNNs in image synthesis applications [125]. Nevertheless, GANs do face some inherent challenges including "non-convergence", "Mode Collapse" and "Vanishing Gradient" [117]. The reader is referred to [117,124] for a detailed discussion of GANs, challenges, and solutions. In the following discussion, we present a narration of the GAN-based synthetic medical data generation. It can be seen from Fig. 7 (see Table B.6, Appendix B for details) that a significant proportion of included studies fall into this category; therefore, it is important to mention that, since this section contains a majority of the included studies, it is longer than others. Thus, for better readability and comprehension, we structure our discussion by breaking it into seven sub-sections (highlighted in bold), according to the seven major facets of GAN development for medical SDG. These are shown in Fig. 8. The initial work

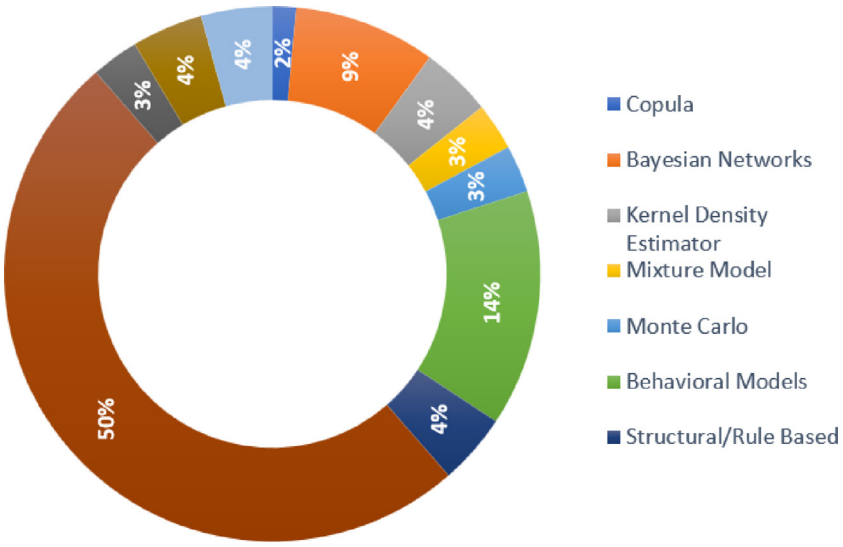

Fig. 7. GMs employed by included studies.

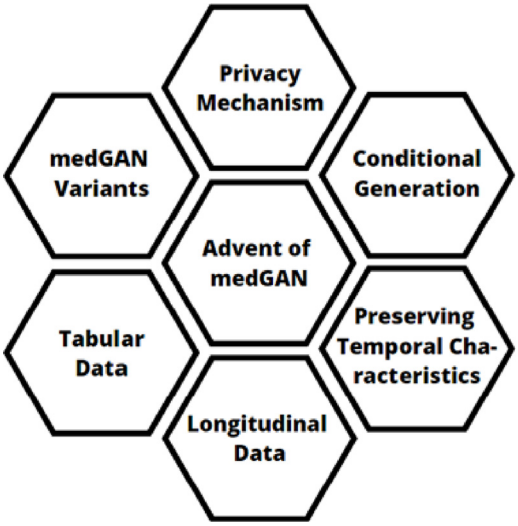

Fig. 8. GAN-based SDG - The development fronts.

included basic models with limited capabilities to spark research in this domain. This is marked by the ***advent of medGAN*** [81], soon after which many ***medGAN-variants*** appeared. However, most of these initial models inherited the limitations of medGAN which primarily included a lack of support for ***temporal and sequential properties*** of data and an inability to generate mixed feature types. Later studies targeted this limitation and proposed other GAN variants to work with mixed features and temporal attributes. Researchers also explored ***conditional generation*** to improve realism. Another direction of research was GANs with ***built-in privacy*** mechanisms. ***Tabular Data***, though similar to snapshots, is being treated as a distinct format that is considered suitable for many data mining applications. Finally, a more recent area of interest is the ***longitudinal synthetic data*** generation using GAN, although, little success has been found in this direction. The next paragraphs narrate the discussion on GAN-based synthetic medical data generation studies and a summary is presented in Table 4.

- **The advent of medGAN:** Use of GAN to synthesize medical data was initially hindered by their inability to work with discrete values [126] which is an integral component of EHR. The first use of GAN for medical data synthesis, named medGAN appeared in 2017 [81] which introduced an auto-encoder into the basic GAN architecture to address this limitation and employed minibatch averaging to combat mode collapse. MedGAN produces an aggregated representation of longitudinal EHRs as binary or counts data in which the rows represent a patient record, and the columns are disease codes. The univariate (dimension-wise statistics) and multivariate (dimension-wise predictions) assessment of synthetic data showed promising results against baseline competitors including Variational Autoencoder (VAE) [127] and Stacked Restricted Boltzmann Machines [128]. Experiments were conducted on 3 different datasets and produced consistent results. Qualitative assessment by medical experts also confirmed general in-distinguishability between synthetic and real records except for a few discrepancies (e.g., both male and female disease codes in a record) resulted because of biases in the original data. Choi et al. [81] argued that privacy is intrinsic to the synthetic data generated through medGANs since no direct mapping between the original and synthetic records exists. The authors backed this claim with results from extensive experimentation performed to assess the privacy leakage in terms of both membership and attribute disclosure bounded against attacker's knowledge. The results confirmed minimal privacy risks, details of which can be found [81]. This work demonstrated a built-in privacy-friendliness of GAN models for synthetic generation. We noted that following these subtle leads, most of the future articles employing medGAN or variants, skipped an explicit privacy evaluation of their approaches assuming inherited privacy guarantees from medGAN. Nevertheless, privacy remains a fundamental concern for medical data sharing and an empirical evaluation of synthetic data for privacy is always desirable before release to public.
- **MedGAN Variants:** The success of medGAN opened a whole new avenue for synthetic medical data generation using different variants of GAN and medGAN. Interested readers are referred to [45] for a survey of GAN-variants employed for synthetic health data. The original medGAN generated discrete data only and this limitation was carried forward in many subsequent models as well. Piper Jackson and Marco Lussetti [129] extended this model in 2019 to include 9

demographic features with different modalities while maintaining the quality of synthetic data, yet again lacking support for continuous data. Jackson's work lacked an empirical re-assessment of privacy for their GAN model under the inherited privacy assurances from medGAN but suggested the inclusion of differential privacy as future work. In an attempt to improve the quality of SD, researchers tried different GAN architectures such as Boundary seeking GAN (BGAN) [130] for medBGAN [86]. BGAN can potentially work with mixed feature types, however, medBGAN is evaluated only with the aggregate discrete data [86]. A sister model medWGAN [87], by the same authors, used a Wasserstein GAN (WGAN) with gradient Penalty (WGAN-GP) [131] inspired by faster convergence and better sample space coverage of WGANs [131]. Experiments showed superior performance of medBGAN against medGAN and medWGAN under all evaluation metrics. The authors asserted inherited privacy assurances from the medGAN and skipped an explicit assessment in these studies. Another project titled Realistic Synthetic Data Generation Method (RSDGM) [85] enhanced the basic medGAN with the inclusion of lab test codes and mixed feature types synthesis. The study assumed privacy inherent in the GAN-based generation. The scores from standard realism metrics showed acceptable realism, However, the training dataset included only a small number of instances and dimensions. RSDGM needs thorough experimentation with full-scale medical datasets for accurate evaluation. Yet another variant of medGAN named MC-MedGAN appeared in [22]. The study evaluated their GAN model against more classical approaches including Independent Marginals (IM), Bayesian networks, mixture model, and Categoric latent Gaussian Process (CGLP). The results showed apparently capricious outcome with the classical approaches outperforming MC-MedGAN in capturing the inter-component relationship of the real data. Nevertheless, the latter demonstrated better resilience against membership and attribute disclosures. This study emphasized the need for deeper analysis of GAN variants to explain and overcome its limitation in producing realistic synthetic data. Some later variants of medGAN such as EMR-WGAN [90] employed significant architectural modifications and are discussed subsequent sections.

- **Preserving Temporal Characteristics:** MedGAN, the pioneering model was an inspiring proposal in the field of medical SDG. The authors of medGAN [81] employed an innovative data transformation that flattened the originally longitudinal ERH into an aggregated snapshot representation which eased the learning. However, this transformation has a major downside as well. Aggregation kills the temporal and sequential properties of the real data which are considered crucial for medical decision-making. Researchers proposed improvements in GAN models to explore alternative models to overcome these limitations. The flattening of longitudinal patient records not only lost the temporal and sequential dependencies but the correlation between different groups of features is also blurred. Yang et al. [43] emphasized the interdependency between disease and medication and proposed Grouped Correlation GAN (GcGAN), designed to learn the correlation between groups of features in a dataset. More specifically, it aimed at capturing the high correlation between a patient's state (diagnosis) and prescribed medication. GcGAN implemented the generator through Fully Connected Network (FCN) to capture medication and its "efficacy" from the data as an implicit representation of disease-drug correlation. A comparison with medGAN and other baseline models proved GcGAN with dense fully connected generator, to be the best-performing model under statistical and utility metrics. The model, however, lacked support for temporal dependencies and continuous data. A similar concept was employed in Sequentially Coupled Generative adversarial Networks (SC-GAN) [39], which incorporated dual generator components to recognize the high correlation between a patient's state (diagnosis) and prescribed medication. Both generators work in conjunction to capture the state-medication interaction to build timeseries records, where for each timestamp, the current state is generated by the first generator and the corresponding medication dosage by the second. The quantitative and qualitative evaluation of generated SD showed promising results compared against the existing models. However, the inter-feature correlations were weaker in the synthetic data. Beginning with medGAN, most of the subsequent models did not learn the temporal and sequential characteristics of the real data, which Torfi et al. [88] attributed to the MLP-based networks in these models. They noted that Convolutional Neural Networks (CNNs) have a better ability to capture the temporal and sequential relationships and used the same as a basis for their model Correlation-Capturing GAN (CorGAN) [88]. The experiments included two datasets: (1) aggregated binary and (2) timeseries data. The results demonstrated the superior performance of CorGAN against the existing models in capturing the correlational structure of the dataset having discrete variables. More significant is the ability of CorGAN to learn the temporal characteristics of the timeseries which most of the earlier models lacked. The authors also established satisfactory resilience in the generated SD against attribute and membership disclosures. Another type of timeseries medical data is the biomedical signals that are an important source of information for practitioners, especially in the critical care setting. Synthetic Signal GAN (SynSigGAN) [89] employed a generative adversarial network for synthetic biomedical signals of various types including electrocardiogram (ECG), electroencephalogram (EEG), electromyography (EMG), and photoplethysmography (PPG). The results show a high correlation between the original and synthetic signals data. However, an explicit empirical evaluation of privacy guarantees was missing in the study [89].
- **Conditional Generation:** The diversity in medical data is a key factor making learning hard for machine learning models, particularly with less training data. Researchers observed that conditioning the generation upon certain features such as phenotype resulted in better convergence and improved realism [90]. A Conditional or CGAN [132] employs a generator that is leveraged to produce data within a certain region of the overall sample space under a given condition. The condition is provided as an input to the generator along with the noise vector, e.g., a generator may be passed a disease code as input to generate synthetic records with that diagnosis. Electronic Medical Record WGAN (EMR-WGAN) and EMR-CWGAN based on WGAN and CGAN respectively, were proposed in [90]. The authors argued that an autoencoder degraded SD realism by injecting additional noise and thus eliminated it from their models. The study [90] presented several noteworthy findings. The authors criticized the existing statistical realism metrics such as Dimension Wise Statistics (DWS) for providing misleadingly optimistic results even when the generated records have substantial discrepancies. They further argued that standard SD utility metrics evaluate its suitability for a particular machine learning task which is a subjective assessment

and may not generalize well to wider applications. The authors [90] proposed new realism and privacy metrics for better quality assessment of SD. The study presented an in-depth comparison of the proposed models against the existing benchmarks and concluded that EMR-WGAN and EMR-CWGAN yielded superior performance scores for the existing and new realism metrics. Nevertheless, the authors [90] emphasized the need for a qualitative assessment of the synthetic records by domain experts as an essential requirement for validation. Like most of other GANs, these models also lacked support for continuous features. HealthGAN [28] is a notable work that adapted medGAN and WGAN-GP for improved stability. Like many other GAN-based models, HealthGAN, in principle, supports continuous features generation only but the use of data transformation techniques produces mixed features in the final synthetic dataset. The study [28] criticized medGAN for its general inability to capture the multivariate structure of the real data. They further noted that univariate realism tests against the individual dimensions provide a superficial assessment of realism and often yield misleadingly optimistic results. This study introduced "nearest neighbor Adversarial Accuracy (AA)" which is an array of metrics to assess realism and privacy together (see Section 5.3). The study also included a workflow for exporting the generative model after training in a secure environment which can enable on-the-fly generation of synthetic data, provided, a small "footprint" of the real data (described as the amount of real data embedded in the generative model) is maintained. A comparison with benchmark models, including medGAN established that HealthGAN-generated SD surpassed the competitors in overall quality. This study [28] made some remarkable contributions regarding the utility of SD for practical applications. These included using SD for classroom teaching and replication of research studies. The authors concluded that, synthetic data maintains adequate utility for educational purposes but pointed out that further experimentation was needed to draw any conclusive outcomes for research and clinical applications. Yale et al. carried out further experimentation to assess the utility and privacy of HealthGAN-generated SD, specifically against membership inference attacks, in [133]. They employed "discriminator testing" to measure the resemblance between real and synthetic records. Discriminator testing is essentially a variant of the distinguishability metric that uses the (trained) GAN discriminator as a classifier to identify between real and synthetic records. The results assured improved privacy preservation for competitive utility against the baseline models. HealthGAN was later employed for synthetic timeseries medical records in [94]. The authors proposed a data transformation function that converted the longitudinal patient record into cross-sectional data without losing temporal properties. The results confirmed the plausibility of HealthGAN for the purpose, however, the experiment was based on a small dataset and the results may not be generalizable for full-blown medical data. Another study [38] employed HealthGAN and generated reasonable quality synthetic data with minimum privacy loss. The authors emphasized a need for further research in this domain to facilitate a steady supply of "re-created" sharable data as a substitute for the classified datasets. A couple of recent studies employed HealthGAN for their experiments with synthetic data. Bhanot et al. [95,96] generated timeseries data using HealthGAN and concluded that the model is unable to capture the time trends of real data accurately. Another study proposed privacy-preserving GAN (pGAN) [93] and employed the adversarial accuracy metric of the HealthGAN. Rashidian et al. put forth the concept of sharpness and smoothness for a conditional generator in the context of their model SMOOTH-GAN [40]. Sharpness refers to the ability of the generator to produce patient states (medications and labs) closely aligned with the underlying disease conditions whereas smoothness allows smooth variations in the generated values, representing the progression of the medical condition. SMOOTH-GAN [40] is based on the CGAN and WGAN-GP architectures, designed to produce both medication and lab measures conditioned on disease codes. SMOOTH-GAN generator can generate "counterfactual or hypothetical" records (by varying the smoothness), which together with explainable AI (XAI) can help gain better insight into disease comorbidities. The first use of GANs for synthetic Gene Expression Data (GED) appeared in [41]. Authors have validated their Gene Expression Generator (GEG) for colon and breast cancer diseases utilizing the publicly available GED data. The synthetic GED data was primarily used to augment the real data for better classifier training. The results showed an improvement in classification accuracy with augmented synthetic data. Another use of conditional GAN appeared in [92]. The model Medical Text GAN (mtGAN) employed an LSTM for the generator and CNN/RNN for the discriminator to synthesize medical text conditioned on disease specification. The study employed a Chinese corpus of clinical notes to train the model. Synthetic text was evaluated using distinguishability metric and augmentation test and the results proved viability of the model. The authors assumed privacy through de-identification of the input, but the generated data may still suffer from attribute disclosure.

- **Privacy Mechanism:** The earlier works in SDG were more focused on realism than privacy, apparently under the perception that synthetic data, being artificial, is intrinsically immune to privacy attacks [39]. However, over time, vulnerabilities of synthetic data emerged, and researchers emphasized that privacy evaluation of SD is essential before release. Following this, many researchers proposed generative models with built-in privacy mechanisms. Differential Privacy has become a De-facto in privacy-preserving machine learning [134] and thus also found its way into the GAN architecture. PATE-GAN [97,135] incorporated the Private Aggregation of Teacher Ensembles (PATE) [136,137] framework for synthetic data with differential privacy assurance. PATE framework embeds differential privacy right into the classifier through an indirect learning mechanism [136]. The idea is to have multiple teacher classifiers which train on private samples and a student discriminator who learns from the aggregated output of the teachers and never gets to see the real private data itself. PATE-GAN employs the PATE framework in the discriminator module by introducing necessary modifications. For details of PATE adaptation to GANs, interested readers are referred to [137,138]. The SD from PATE-GAN is evaluated in three different settings with multiple datasets and superior performance was observed against the benchmark model [139] for various utility metrics. Another attempt at building differential privacy into the GAN architecture named Privacy Preserving GAN (PP-GAN) is presented in [98] which employs a WGAN with controlled addition of noise to the discriminator gradient for DP guarantees. However, the study lacked detailed experimentation results for EHR data. Following the trend, many authors considered differential privacy as a core privacy metric for synthetic medical data. Auxiliary Classifier Generative Adversarial Network (AC-GAN) [99] attempted

to generate synthetic participants for the Systolic Blood Pressure Intervention Trial (SPRINT) with differential privacy guarantees. AC-GANs showed promising results in the preservation of the statistical structure of the original data and credible predictive performances of models trained and tested on synthetic data. The authors also demonstrated a significant correlation between feature ranking and parameter values of the models trained on synthetic and real datasets. Differential privacy provides satisfactory privacy guarantees but comes at a cost, which, in the case of GANs is the additional noise added to the generation mechanism resulting in compromised realism. Wang et al. attempted to address this gap in their model Privacy Preserving Augmentation and Releasing scheme for Time series data via GAN (PART-GAN) [100], which is an enhancement of the AC-GAN. PART-GAN employs conditional generation and some novel optimization strategies which quantify and control the injection of (unnecessary) noise under the DP setting and improve the overall stability and convergence of the generator training. Synthetic data from PART-GAN preserved the correlational structure of the real data while maintaining higher disclosure protection. A recent study [101] proposed a new privacy scheme known as *identity masking* in their model Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). This model is based on WGAN with gradient penalty (WGAN-GP) and CGAN and introduced the notion of "identifiability" in the loss function. Identifiability measures the proximity between real and synthetic records using the standard distance metrics. The authors note that a reasonable minimum distance must be maintained between synthetic and real records to guard against the membership and attribute disclosures embedded in the identifiability constraint into the loss function of the generative model to ensure sufficient separation. Experimental results showed superior performance of ADS-GAN over PATE-GAN [97] and other benchmark models [139] in preserving statistical and correlational structures under privacy constraints. Moreover, the utility evaluation of SD from ADS-GAN showed comparable predictive performance against real data for downstream machine-learning tasks. Imtiaz et al. exploited the post-processing theorem of DP for a simplistic privacy solution [102]. The authors applied DP to the input data instead of the generative model. The post-processing theorem suggests that a function applied to a DP mechanism will also satisfy the same DP guarantees [102]. The study [102] utilized a BGAN to generate aggregated synthetic timeseries data. The real data consisted of daily food and activity logs from 25 individuals, collected privately using the Fitbit smartwatches and API. The synthetic data showed good univariate similarity with the real data, however, a deeper analysis of the multivariate metrics was missing in the study. A recent study employed Renyi DP, a relaxed variant of traditional DP, in [104] for the model Renyi DP Convolutional GAN (RDP-CGAN). The study also included a few other modifications in the generic medGAN architecture. Firstly, convolutional neural networks were used for all components i.e., generator, discriminator, and autoencoder. Second 1-D convolutions were utilized to capture micro-level correlations. Third, Renyi DP was enforced under a privacy accountant to calculate the privacy loss. The authors of [104] also performed an ablation study and concluded that convolutional architecture performed significantly better than regular NNs.

- **Tabular Data:** Tabular Data is similar to the snapshot data (as defined in this study) which represents one small part of a patient EHR. Although most of the SDG methods end up generating data that is presented as tables, more recently, researchers emphasized the term "tabular data" to highlight unique challenges associated characteristically with such structures. Tabular data typically contain mixed feature types and may suffer from the curse of modality (multimodality) and unbalanced frequency of values in individual columns. Park et al. proposed tableGAN [83] as an enhancement of the deep convolutional GAN (DCGAN), by incorporating an additional classifier in the vanilla GAN architecture. The tableGAN also adds additional losses i.e. information and classification loss, to the general GAN loss, for improved realism. The information loss emulates the DWS metric aiming to reduce the feature-wise distances between the real and synthetic data. Similarly, the classification loss is a variation of the DWP metric to improve the "semantic integrity" of the synthetic data. Evaluation of the synthetic data from tableGAN showed comparable utility for higher privacy against anonymized data. However, a comparison against contemporary GAN models was missing in the study. Fang et al. improved upon Conditional Tabular GAN (CTGAN) [140] to provide differential privacy for synthetic tabular medical data in their model DP-CTGAN [84]. The base model CTGAN takes into account the specific characteristics of the tabular data mentioned above to generate SD that preserves the feature correlations, multi-modality, and uneven distribution of categoric values [140]. DP-CTGAN-generated SD showed superior utility for classification tasks against contemporary models but is not suitable for practical applications in its present form. A recent study [91] stressed that better record-level fidelity can be achieved by ensuring standard constraints between feature values. For example, a record with a gender value male cannot have female-specific disease codes. While many studies included a post-hoc evaluation for inter-feature dependency preservation in the tabular SD (Table 7), authors of [91] penalized the generator in their model Heterogeneous GAN (HGAN) for dependency violations in the generated data. HGAN adapts the EMR-CWGAN [90] by incorporating additional non-linear filtering (ReLU) before conditional normalization layers in the generator and discriminator networks. The authors conclude that HGAN outperforms the parent model EMR-CWGAN in terms of utility while suffering no greater privacy risk.
- **Longitudinal Data:** Most of the early research aiming at producing realistic synthetic EHRs using GANs in particular, focused only on the aggregate, cross-sectional, or snapshot data [45]. However, longitudinal EHRs offer more research value and better innovation potential but have been studied little [34]. Some recent studies [103] emphasize the superiority of longitudinal patient data and employ adversarial networks in this regard. A dual discriminator architecture known as Dual Adversarial Auto Encoder (DAAE) is proposed in [103] to facilitate differentially private longitudinal synthetic EHR generation. DAAE synthesizes set-valued sequences with the help of a recurrent autoencoder working in conjunction with two discriminator networks. The generator aims to match the synthetic data distribution with the encoded distribution in the latent space. The evaluation of DAAE shows superior performance against the baseline models (medGAN) in predictive modeling. Qualitative assessment by medical experts further confirms the superiority of this approach. Another attempt at longitudinal data synthesis is presented in [105]. The study proposes a CWGAN-based model called SynTEG (Synthetic Timeseries Generation) for synthetic EHR of multiple visits. SynTEG follows a two-stage learning approach. The first stage called *dependency learning* estimates the timestamp and the expected patient state (diagnosis to appear given the current

state) of the next visit, which is then passed as a condition to the generator in the second learning stage called *conditional simulation*. The generator produces synthetic longitudinal patient records conditioned on the expected patient state. SynTEG uses an RNN in the dependency learning stage along with auxiliary classifiers and regressors to estimate the next patient state given the previous visits. The second stage adapts the EMR-WGAN [90] for a synthetic generation. The authors [105] utilized SD quality metrics from their previous work [90] with promising results. SynTEG-generated SD showed satisfactory resilience against membership inference attacks. A recent study [106] by the same researchers highlighted that the synthetic distribution tends to drift away from the real for longitudinal simulation because of exposure bias, a phenomenon common in sequential generative models [141], and propose an automated pipeline for the generation of realistic synthetic longitudinal data. The study employs Conditional Fuzzing and Regularization (CFR) to mitigate the accumulation of error in the sequential generation which causes the drift. Regularization is achieved by maximizing the mutual information between consecutive generated sequences of visits (called episodes) while a controlled noise fuzzies real sequences exposed to the generator during training. The authors demonstrate a substantial decrease in the drift with CFR but urge further investigation for longer sequences. The study lacked an empirical assessment of privacy.

2. Neural Network

The second generative model that we discuss for DD methods is Neural Network (NN). Neural Networks are mathematical models that mimic the human brain for the storage and processing of information and can model complex patterns. NNs consist of thousands of interconnected computational units called neurons which work collaboratively to learn highly complex non-linear patterns. Neural Networks and their variants (RNN, CNN, etc.) have been used extensively in the healthcare domain for diagnostics, disease predictions, and image segmentation to name a few [142]. Inspired by their ability to self-learn highly complex structures, researchers have employed neural networks for synthetic health data generation. It should be noted that they also form the foundation building blocks for almost all of the ML-based generative models including GANs. Nevertheless, NNs and their variants such as Recurrent Neural Networks (RNNs), have also been used as standalone generative models in various studies. RNNs are a special variant of NNs, capable of learning temporal dependencies, and have been employed for bio-medical signal generation [107]. The study used a bidirectional recurrent neural network (BiRNN) to generate partially synthetic biomedical signals of 5 different types including ECG, EEG, BCG, PPG, and Respiratory Impedance. The quality of the synthetic signal is evaluated based on the disagreement with the real signal at sampled points. Classifiers trained to differentiate between different types of ECG signals using synthetic data can achieve an accuracy of up to 99%. The study [107] left a detailed evaluation of privacy to a future publication.

3. Autoencoders (AE)

An autoencoder is a neural network that attempts to learn an efficient representation of the input data. In addition to being used as a core component (to support the generation of discrete data) in various GAN-based SDG architectures, autoencoders have also been utilized as standalone generative models. Medical data is both high-dimensional and highly inter-correlated. Earlier research has established the success of marginal distributions in capturing the underlying structure of the data [77]. Abay et al. [108] employed a deep learning approach to learn the marginal distributions of classification datasets through AEs for each class cluster in the data. The autoencoder learns the distribution of each cluster which is further passed through the Differential Privacy Expectation–Maximization (DP-EM) function for noise injection under the differential privacy criteria. Synthetic data is then sampled from this private latent distribution. Statistical comparisons between original and synthetic data as well as utility evaluation of SD showed promising results and authors claimed improvement over a popular competitor PrivBayes [75]. However, the method strictly needs labeled input data which is a challenging requirement in case of already scarce medical data. Moreover, the generated SD is good for classification tasks and lacks generalization for other applications. A variation of autoencoders termed adversarial autoencoders have also been used for synthetic medical generation in [103]. Electronic health record Variational Autoencoder (EVA) [109] targets longitudinal EHR generation by employing an autoencoder as the generative model. The study also presents a conditional variant of their model called EVAc, capable of generating records with specific diagnoses, provided as input to the generator. The model splits the overall learning as patient-wise and population-wide optimization running in parallel for improved local (instance) and global (dataset) realism. The study [109] presented results from a thorough evaluation of the generated SD through several qualitative and quantitative metrics shown in Table 4. The authors also performed a membership disclosure assessment and concluded that the generated SD provides sufficient protection against such attacks.

4. Decision Trees (DTs)

Decision trees are among the simplest machine learning model with human interpretable structures. Researchers regard interpretability as an important characteristic of ML models which not only increases trust but also provides actionable insight towards prospective improvements. Interpretability is one dimension of the recent drive on Explainable AI and is highly desirable in the medical domain owing to the impact of medical decisions. Interpretable models provide deeper insights into the data and open room for improvement. Randomized Decision Trees (RDTs) were employed in [31] to generate synthetic data that demonstrated high-performance scores for classification and regression tasks when tested on held-out real data. The predictive models trained on the synthetic and real data showed similar feature ranking which is indicative of high resemblance. The authors asserted intrinsic privacy guarantees inherited from the RDT structure which incorporates differential privacy requirements by default. A well-cited work by Emam et al. [110] studied the effect of variable order in sequential decision trees for data synthesis. The authors employ a simple sequential generation strategy by varying the order of the variables/features. The process runs in multiple cycles with one feature generated in each cycle completed in two steps. The first step builds a tree for the selected feature followed by a synthetic sample drawn from this tree in the second step. Each subsequent feature tree is conditioned on the previously generated features. The authors [110] concluded that the order of feature generation impacts the quality of synthetic data. The univariate realism was evaluated using the Hellinger distance between real and synthetic data distributions. Dimension-wise average and discriminative score metrics were employed for multivariate realism. Moreover, the utility analysis of synthetic data produced satisfactory accuracy scores for hospital re-admission and treatment arm prediction with synthetic data. A later study, by the same authors, applied the approach for synthetic COVID-19 dataset in [111] with a specific goal to assess the plausibility of this data as a proxy for real sensitive data to support analysis. The study [111] performed a comprehensive evaluation of the

generated SD utilizing various quality metrics and concluded that this dataset provides sufficient realism and privacy to be shared for analysis and research purposes.

5.2.3. Hybrid data generation

Hybrid techniques build generative models with truth extracted from data as well as theory. This kind of generation particularly suits medical data because of the high realism requirements [143]. The DD part enables learning of the basic overall structure of the real data and the knowledge from theory can furnish it with specific details for better realism. In the following paragraphs, we discuss the hybrid SDG studies categorized according to the generative models employed. These categories include: (1) Behavioral Models, (2) Structural (3) Bayesian Networks, (4) GANs and (5) Neural Networks. Table 6 presents a summary of the methods with a focus on knowledge and data sources, generative model, SD granularity, and features supported. We also highlight the strengths and limitations of each.

A. Behavioral Models

One of the notable hybrid data generation works is presented by Buczak et al. [24]. Longitudinal EMRs are developed as a three-step process beginning with generating synthetic patients (demographics and chief complaint) as a data-driven task by learning the health incidence and disease prevalence from real EMR. Next, care patterns are derived from similar real records including diagnoses, labs, prescriptions, reports, and alike. The final step included a KD-based adaptation of the care map which is then attached to the synthetic record. Care is taken during the care map adaptation to avoid complete replication of a real record. The study [24] generates synthetic data to model the outbreak of the tularemia fever epidemic and generates two types of synthetic patients; (1) day-to-day patients with various diagnoses (as per the incidence) and (2) others injected with tularemia symptoms. An expert inspection of the generated records showed sufficient realism however, manual adjustments to the synthetic care model were needed in some cases to remove discrepancies or errors. An interesting idea of a hybrid between synthetic patients generated in knowledge-driven mode from Synthea and DD timeseries information from MIMIC is presented in [145]. Synthetic patients with demographics and basic disease information are obtained from Synthea which are then matched (against as many attributes as possible) with the records in MIMIC. The timeseries information from the matched records is extracted to furnish the synthetic records. A key strength of this approach is the simplicity of implementation, but the generated data is prone to attribute disclosure attacks. The study [145] presented only a superficial summary of their experiment; however, the proposed idea can be extended for more advanced data synthesis. Larrea et al. carried forward the same approach [145] for their method in [146] by employing a python-based SDG library Synthetic Data Vault (SDV) [151]. The study [146] generated partially synthetic data in four main steps; (1) generation of synthetic subjects from meta data (2) Extraction of important statistics from real-time series data and appending it to the subject metadata, (3) feeding the metadata to SDV (or in principle to any other SDG) to generate synthetic records (4) closest matching real-time series is then added to the synthetic record. The simplicity of the method is a plus; however, the partially synthetic nature of the generated data makes it vulnerable to privacy leaks. Another method for synthetic timeseries data generation appeared in [27] which employs Facebook's time series prediction API "Prophet" to synthesize diabetic foot data. Prophet is an open-source forecasting library that allows modeling periodic and a periodic trend as well as holidays and errors. The study demonstrated an intelligent use of this model to synthesize timeseries data for diabetic foot. The ground truth was extracted from the UCI Diabetes dataset as well as knowledge from medical research to construct the desired training dataset. The synthetic timeseries was sufficiently close to the real training data. An extension of this work appeared in [29] which employed an advanced version of the Prophet library called *NeuralProphet* model for improved realism in capturing various trends in the timeseries data.

B. Structural Methods

A rule-based hybrid approach for synthetic longitudinal EHR generation is presented in [144]. The method works in two phases. The first phase extracts prevalence information from real data to generate virtual patients' demographics and basic information. The second phase simulates the disease in virtual patients under the constraints/rules recorded in decision tables. Decision tables are manually curated from the disease progression models and associated CGPs. Each step, termed an "episode" in the EHR, is an outcome of a health event on the current state of the patient which can lead to improvement, degradation, or no change in the health conditions. Multiple episodes are generated to build complete EHRs. A major strength of this work is its simplicity; however, substantial manual effort is required to build decision tables with an appropriate level of detail. Synthetic clinical notes have been studied less in the literature so far [148]. A text generation method through expert-curated disease models as well as statistical information extracted from real data is presented in [35]. The real data serves as a source for personalization in the synthetic text, representing different writing styles and errors present in the real environment. The authors evaluated their method by generating synthetic clinical notes for mental illness. The initial results are promising. However, in its present scope, the method is heavily knowledge driven and thus demands more human intervention.

C. Bayesian Networks

Zhenchen Wang et al. [143] proposed a hybrid method using BNs as a three-stage process including: (1) ground truth selection, (2) data generation, and (3) evaluation. Ground truth selection includes the amalgamation of knowledge acquired from real data and domain theory to build (real data) and refine (domain knowledge) an 8BN. Synthetic records are then extracted from the BN by random sampling in the second phase. The authors evaluated the synthetic data for univariate and multivariate realism through distance-based realism metrics and pairwise correlation matrix respectively and claimed a satisfactory resemblance between real and synthetic distributions. The NMDS plots of real and synthetic datasets clusters showed considerable harmony indicating good realism. For privacy, the authors employed fundamental checks including non-identifiability assurance (synthetic instances are not too similar to real records) and sufficient separation between synthetic and real data outliers. Tucker et al. [21] emphasize that biases and missing information in the real data deteriorate the quality of SD generated from a model trained on this data. They advocate transparency and explainability of the generative model so that inconsistencies and discrepancies in the generative model may be traced and corrected by domain experts. The study [21] further noted that because of the inherent complexity of medical data, many latent relationships exist within the features and an explicit representation of the same can improve SD realism. The method followed a three-step process; (1) A BN was extracted from the ground truth (2) the BN was annotated with the "data missingness" information by incorporating additional nodes. (3) Latent variables and relationships were identified and represented explicitly in the BN. Synthetic Data was then sampled from this fine-tuned BN. The quality evaluation of SD showed significant improvement in statistical realism. Outlier similarity and reproduction rate metrics were employed to assess privacy. However, the detailed results of the evaluation were missing

Table 6
HBD synthetic medical data generation methods (Feature Type: M: Mixed, C: Continuous, D: Discrete | Synthetic Dataset Format- L:Longitudinal, S: Snapshot, A: Aggregate, T: Temporal).

Study	Knowledge source/ Real data	GM	SD granularity	Feature types	Realism validation metrics	Privacy metrics	Strengths	Limitations
[24]	Academic and Research Literature, Domain Experts/ CDC NSSC Biosense data	Computer Programs/ Simulation Scripts	L	M	Expert Inspection	Reproduction/ Duplication Ratio	Supports longitudinal data generation. Datasets with specific characteristics can be generated with a relatively small dependence on human experts. The process is mostly automated.	Depending on human intervention for care pattern adaptation. Poor scalability for large dataset generation.
[144]	Clinical Practice Guidelines, Domain Expert/Real Private EMR	Decision Tables	L	M	Expert Inspection	Default	The simplicity of modeling of GM. Supports longitudinal data synthesis.	Multimorbidity is not supported.
[35]	Medical Literature, Domain Expert/MIMICC	Manually Curated Templates	Text	–	Expert Inspection BLEU, METEOR	Default	Simplistic model for clinical notes synthesis. Captures writing styles and errors of the real world.	Too much human effort is needed in the current version.
[143]	Domain Expert/UK primary care data	Bayesian Networks	S	M	Dimension Wise Statistics KS-Test Visual Plausibility: NDMS, Correlation heatmap	Outlier Similarity	Simplicity of GM.	Lacks support for longitudinal records.
[145]	Synthea/MIMIC	Behavioral Model Synthea	T	M	–	Default	The simplicity of the approach.	High privacy risk
[146]	Meta Statistics/TMET	Behavioral Model/SDV	T	M	Moment Similarity Visual Plausibility: Correlation Heat Maps,PCA plots	Nearest Neighbor Metric	The simplicity of the approach.	High privacy risk.
[27]	Domain Expert, Medical Literature/UCI Diabetes Dataset	Behavioral Model -Prophet Library/ NeuralProphet Library	T	C	Pair-wise Correlation Difference Visual Plausibility – Histogram Clustering Class Satisfaction Accuracy	Inherited	Timeseries can catch multiple parallel trends of different granularity i.e., daily, weekly etc.	Suited to timeseries data only.
[21]	Domain Experts/CPRD Aurum Dataset	Bayesian Networks	S	M	Dimension Wise Statistics Feature-wise Distributional Distance KL Divergence Kernel MMD Performance Agreement	Outlier Similarity Reproduction rate	Minimizes the effects of real data biases. Explainable generative model.	Poor scalability for high-dimensional datasets. Lacks support for longitudinal EHR.
[147]	Domain Experts, Medical Literature/MIMIC, CPRD Aurum Dataset	Bayesian Networks, Dynamic Bayesian Networks	S (MIMIC) T (CPRD Aurum)	D C (M))	Visual Plausibility – Correlation heatmap, histograms KL Divergence	Outlier Similarity Nearest Neighbor metric	Captures temporal features of real data.	Poor scalability for high dimensional dataset. Lacks support for longitudinal EHR.
[49]	Domain Expert, Medical Literature/MIMIC, Nephrotic Syndrome Dataset	HA-GAN	A	M	Performance Agreement	Default	Can learn from a smaller dataset. Comparatively lesser human effort is required. Robust against training dataset imbalance.	Lacks support for longitudinal data. Not scalable to a high-dimensional dataset.
[148]	Domain Experts/Private Clinical Text	LSTM GPT-2	Text	–	Performance Agreement – Classification Expert Inspection Augmentation Test Clinical Applications -De-identification	ROUGE-N Expert Inspection	–	Generated data contains high disclosure risk for sharing.

in the study. The same authors extended their work for time-series data synthesis in [147] by using a Dynamic Bayesian Network (DBN). However, the experimentation was performed on a small and specialized dataset with a focus on only two temporal features namely Systolic and Diastolic Blood Pressure (SBP, DBP). The authors used Structural Expectation–Maximization (SEM) to learn a DBN from the data, which was then manually augmented with additional known relations. Samples drawn from this model

Table 7
Resemblance metrics used in included studies.

Resemblance metric classes	Metrics	Studies employing the metric
Univariate	Dimension Wise Statistics	Moment based [39,40,43,73,74,78–81,86–88,91,101,143,146] Support Coverage [22] Directional Symmetry [95,149] Rare Feature Occurrence Rate [80] Comparison against Reference/Real Statistics [23,25,32,71,72,78,79,111]
	Distance Based	MSE, MAE [77,89] PRD,RMSE [95,107] KS [21,23,42,80,85–87,102,143] Kullback–Leibler Divergence [21,22,147] Fréchet Distance [89] Hellinger Distance [110] Inception Score [100] Short timeseries Distance [95,149] Pearson Chi-squared Test [23] Wilcoxon/Mann–Whitney Rank-Sum Test [23]
Multivariate	Dimension Wise Prediction	[22,75,80,81,86–88,90,91,104,104,110]
	Correlation/Covariance Based	Pair Wise Correlation Coefficient [22,23,77,80,89,95,99,107,143,147] Feature Co-occurrence [43] Cross-Type Conditional Distribution Statistics [91]
	Association Rule Preservation	Association Rule Mining [80,86,87,91] Domain Rule Preservation [22,32,91]
	Indistinguishability/ Discriminator Metric/ Distinguishability Score	Propensity Score [73,74] Discriminator Metric [80,92,103,106,110,111] Discriminator Testing [133]
	Latent Space Metrics	Log Cluster [22] Latent Space Representation [90,91,105]
	Distance Based	First Order Proximity [90,105] Euclidian Distance [143] Jenson -Shannon Divergence [101] Wasserstein Distance [101] Total Variation Distance [75,108] Maximum Mean Discrepancy (MMD) [21,104] Inception Score [100] Jaccard Similarity Index [150]
	Combos	Adversarial Accuracy [28,38,93] Identifiability loss [101]

showed substantial univariate and feature correlation similarity with the original data. The Authors of [147] also performed an outlier similarity test to mitigate the privacy leakage.

D. Generative Adversarial Networks

Human Allied GAN (HA-GAN) [49] present a different take on hybrid data synthesis by incorporating a mechanism of "advice" to the GAN. The authors employ a WGAN annotated with human guidance at fixed intervals during training. The idea is to inject "expert advice" into the generation mechanism after a certain number of iterations of regular training to speed up the convergence as well as tune the generation. This is particularly useful in situations when the volume of real training data is low. The study [49] uses a feature correlation matrix as the injected advice, the effect of which reached the generator through the discriminator loss function. Not only was the model able to learn from a small dataset (50 records) but also showed resilience against imbalance in the training data. The utility evaluation of SD from HA-GAN showed improved results as compared to benchmark models including medGAN [81], medBGAN [86], and medWGAN [87].

E. Neural Networks

Long Short-Term Memory (LSTM) is a special type of RNN that can learn long-term dependencies from sequential data. LSTMs have been used extensively in NLP. Libbi et al. Study [148] employed an LSTM to produce synthetic clinical text with a specific aim to make privacy-safe data available to train machine learning models for Name Entity Recognition (NER) downstream tasks (de-identification in this case). The author annotated personal health indicators (PHIs) in the real data and trained a generative model to produce similar documents. The method generated partially synthetic annotated text that was used to train a machine learning model to perform automatic de-identification of PHIs. The results showed satisfactory performance for the de-identification of PHIs. However, with larger notes, big parts of the real text were copied into the synthetic data. The study [148] included results from an experiment with GPT-2 as a generative model but the authors concluded that LSTM-generated data offered better utility for the de-identification task.

5.3. Quality evaluation of synthetic data

In this section, we discuss our findings regarding the quality evaluation of SD to answer our RQ4. Synthetic datasets are required to exhibit two quality attributes to a reasonable proxy for real data. These include (1) Realism and (2) Privacy. Realistic synthetic datasets possess the same statistical characteristics as the real data but no real values and hence privacy laws are not applicable. Nevertheless, synthetic datasets do face certain disclosure risks, which must be carefully assessed before sharing. Fig. 9 depicts the important concepts regarding the quality assessment of SD. In sections 5.3.1 and 5.3.2 we discuss the realism and privacy evaluation respectively.

5.3.1. Realism validation

Realism can be defined as the quality of synthetic data being "sufficient to replace real data" [69]. The higher the realism in

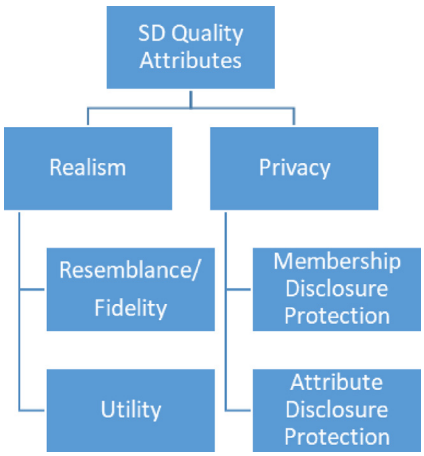

Fig. 9. SD Quality Evaluation.

synthetic data the better they can substitute the original in practical applications. Realism has two dimensions; (1) Resemblance or Fidelity and (2) Utility. The former defines how close the synthetic dataset is to the real one in statistical characteristics such as feature distributions and correlations. The utility is simply the usefulness of synthetic data for a practical application. We classify the realism metrics based on the mode of assessment as either ***Quantitative*** or ***Qualitative*** and discuss these in the following subsections.

A. Quantitative Assessment of Realism

The quantitative realism evaluation is based on the structural and behavioral assessment of similarity between the synthetic and real datasets measured as a numeric value. Although little standardization has been done for SD quality evaluation [152], the quantitative assessment metrics, provide common grounds for comparison of methods, as they are mostly based on the standard statistical measures of similarity between data distributions. We discuss the quantitative metrics grouped as Resemblance/Fidelity and Utility. Fig. 10 presents a classification of quantitative SD evaluation metrics.

1. **Resemblance Metrics:** Resemblance is essentially a basic sanity check for the synthetic data before a more rigorous quality assessment. These metrics employ some standard measures of similarity between the univariate, multivariate, and correlational structure of real and synthetic data. Table 7 lists the different resemblance measures used by researchers in this regard. ***Univariate metrics*** attempt to measure the similarity between the distributions of individual dimensions in real and synthetic datasets. *Dimension Wise Statistics (DWS) is* the most widely used univariate metric. Researchers have employed moment-based comparisons e.g., comparing the means, proportions, or higher-order moments of corresponding features in real and synthetic datasets. In addition to these, the support coverage metric measures the coherence among the frequency distributions of synthetic and real categoric features A variant of this is the rare feature occurrence metric that compares the proportion of rare values in synthetic and real features. For KD approaches, since no real dataset is involved, synthetic feature statistics are compared against known reference statistics such as health incidence reports. Statistical measures of univariate distribution distance have also been used in the literature to assess the (dis) similarity between corresponding feature distributions of real and synthetic data. Kolmogorov Smirnov (KS), Jenson Shannon Divergence (JSD), Fréchet Distance, and Hellinger, to name a few, have been used for the purpose. The univariate (dis) similarity values for each dimension of the synthetic dataset can be aggregated (for example through MSE) to get a cumulative univariate (dis) similarity score. Bhanot et al. [96] noted that regular metrics are not suitable for the evaluation of timeseries data and proposed Directional Symmetry (DS) and Short Timeseries Distance (STS). Both metrics target the covariates in the data to assess the similarity between the trends depicted by real and synthetic series. DS measures the symmetry or harmony of the up or downward direction of trends between the real and synthetic series while STS approximates the average distance between the two. In addition to individual feature similarity, inter-dimensional relationships are of utmost importance in any data in general and EHR in particular [16]. The multivariate structure of the synthetic and real datasets must depict high similarity for the former to be of practical use [54]. For example, while certain diseases such as cold and fever have a high co-occurrence rate, certain other disease codes e.g., prostrate and uterus cancer cannot possibly be comorbid. Such dependencies cannot be ensured through univariate metrics and it has been shown that a high score on univariate resemblance does not guarantee sufficient multivariate similarity [22]. ***Multivariate Metrics*** are an essential second level of resemblance assessment. *Dimension Wise Prediction* (DWP) is the most employed metric specifically for aggregate synthetic data (Table 7). It measures the extent to which the inter-dimensional dependencies have been captured in the synthetic data by iteratively taking each dimension of the dataset as a target to be predicted by the remaining dimensions. *Multivariate Distributional Distance Metrics* such as Wasserstein Distance, and Jenson Shannon Divergence, to name a few, have also been used in multiple studies (Table 7). *Correlation/Covariance Based Metrics* use a measure of correlation as a basis to assess the similarity of inter-feature dependency preservation in synthetic data. These include *Pair-Wise Correlation* differences between corresponding pairs of synthetic and real features. *Feature Co-occurrence* measures the similarity between the co-occurrence of discrete feature pairs in real and synthetic data. Multivariate analysis methods, including, *Principal Component Analysis (PCA)* and Non-metric Multi-Dimensional Scaling (NMDS) have also been used for multivariate similarity scores. Another popular multivariate metric is the *Discriminator Metric or Distinguishability* Score which employs a post-hoc machine learning-based classification of synthetic and real records in a supervised mode. It works by training a classifier upon a dataset containing labeled instances of real and synthetic records. An accuracy of 0.5 by this classifier on the test set indicates complete indistinguishability between real and synthetic records further indicating high multivariate similarity. *Association Rules Preservation* measures the inter-dimensional correlation by comparing the association rules mined from real and synthetic dataset. Higher percentage of matching rules indicates higher multivariate realism. A KD variant of this metric is *domain rule preservation* which measures the proportion of known feature associations preserved in the synthetic data. The idea of *Latent spaces* has also been used for multivariate realism assessment in a few studies (Table 7). *Latent Space Representations* (LSR) captures the most significant correlational aspects of the originally high-dimensional data by projecting it to a low dimensional

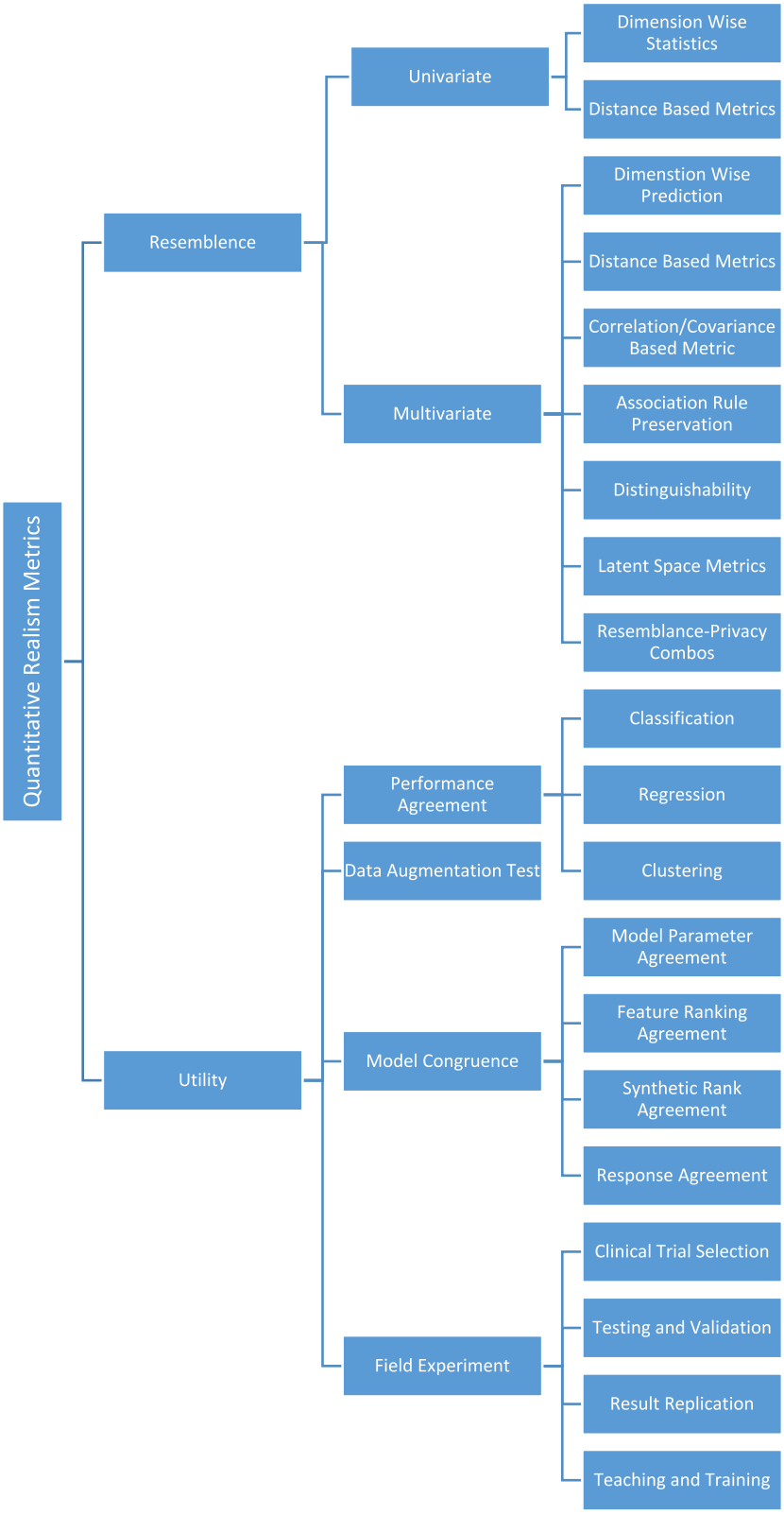

Fig. 10. High Level Classification of quantitative realism metrics.

space for comparisons. *Log-cluster* is another latent space metric measuring the coherence between clustering patterns of real and synthetic datasets. In addition to these, some combination metrics for utility and privacy have been devised by researchers (Table 7). Adversarial Accuracy is a distance-based metric in principle but combines the measure of utility, in terms of distinguishability, and privacy in terms of nearness to real records. Similarly, the identifiability, also discussed in a later section, combines utility and privacy. The fundamental principle behind combos is

to find an optimal balance between realism (go as near the real distribution as possible) and privacy (do not go so close to reveal information) as an integrated objective.

2. **Utility Metrics:** The resemblance metrics above assess the structural or static aspects of the synthetic data for a generic realism assessment. It has been shown that synthetic data undergo lesser loss of utility for machine learning tasks [54] and may perform high despite an average score on fidelity. However, since utility is a subjective matter, the SD may show varied usefulness for different applications in practical settings, and it is important to make a careful judgment of utility. The utility metrics take a pragmatic approach by deploying the SD to the target application and measuring their usefulness. We present a classification of utility metrics in Fig. 10. ***Performance Agreement***, metrics compare the aggregate scores of synthetic and real datasets when applied to the same practical application using relevant performance metrics. Most popular performance agreement metrics employ one instance of ML model trained and tested on real data and the other on synthetic data respectively and comparing the scores. Several other settings have also been used for these metrics including, *Train on real and Test on Synthetic (TRTS)* [153] and *Train on Synthetic data and Test on Real (TSTR)* [153]. Any standard predictive performance metrics e.g., F1 Score can be employed for performance comparison. The closer the performance scores between synthetic and real data sets, the better the realism of SD. *Response Agreement* is a micro-level variant of performance agreement that compares the response of ML agents trained on synthetic and real data for the same test example. Higher realism will result in a higher percentage of matched responses. *Synthetic Rank Agreement* [154] provides an indirect measure of realism. It works by training the same set of classifiers independently on synthetic and real data. Realism in SD should result in a similar (performance-based) ranking of the classifiers in each set. We define ***Model Congruence,*** as an array of white-box metrics, **w**hich look for similarities in the internal configurations of the models trained on synthetic and real data. These include *Parameter* and *Feature Rank Agreement* metrics which respectively look for correspondence in the parameter values of both models and the feature rank order. ***Augmentation test*** metric works by augmenting the real data with synthetic instances to train an ML model. If the performance of this model is better on the augmented set than with real-only training data, it indicates a higher utility of SD. In addition to laboratory evaluations, some authors have employed synthetic data for practical clinical tasks such as clinical trial selection [68] and as test data for clinical support systems [36] (See Table 8). Another study [28] evaluated the efficacy of synthetic data to replicate earlier research studies to see how well the outcomes match. The same study utilized synthetic data in the classroom as part of a data science challenge for UG health informatics students. Subjective quantitative metrics are required to quantify the scores for this kind of pragmatic assessment. Collectively we group them as field experiments.

B. Qualitative Assessment of Realism

Qualitative assessment of synthetic data involves human evaluation of the plausibility of synthetic records. Researchers regard this as a necessary step after quantitative assessment to identify any discrepancies [90]. Qualitative assessment is generally a manual process that largely depends upon the skill and experience of evaluators. We classify qualitative metrics into two categories: **(1) Expert Inspection and (2) Visual Plausibility.** Fig. 11 presents a high-level classification of qualitative assessment methods.

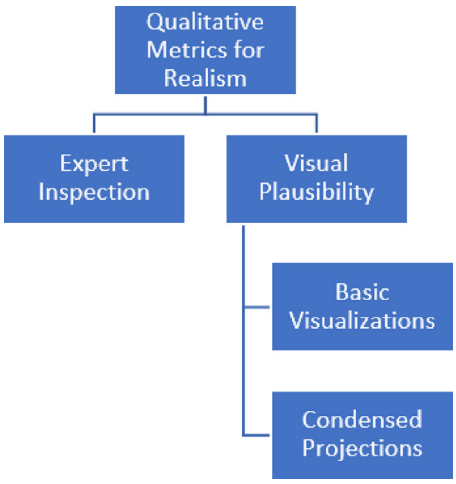

Fig. 11. High-level classification of qualitative metrics of realism.

1. **Expert Inspection** is the qualitative assessment of the synthetic data by the domain experts, the medical practitioner, in the case of medical data. In this process, the domain experts manually examine each synthetic record for its clinical plausibility in terms of adherence to constraints, value ranges, feature correlation, and frequencies. Most of the studies emphasize the need for a final qualitative assessment after initial quantitative evaluation to ascertain the validity and sufficient diversity of the synthetic records. However, this includes a few challenges.

    1. 1. It is a tedious and slow process and human errors might still seep through.
    2. 2. Exhaustive evaluation becomes impractical with large volumes and dimensions.
    3. 3. The process is subjective to the expertise, skill, and commitment of the examiner, and thus different results are expected if different inspectors evaluate the same set of records.

2. **Visual Plausibility** includes data visualization-based comparisons between real and synthetic data to see how well they correspond as an indicator of realism. Several data visualization methods ranging from a histogram to correlational heat maps to PCA have been used in the literature (Table 9). We classify them into two broader categories: (1) Basic Statistical Methods including histograms, heatmaps, and probability curves, and (2) Condensed Projections including dimensionality reduction procedures such as PCA and NMDS. Practitioners can use an array of many other visualization methods depending upon the data formal and realism requirements. The visual assessment not only provides a quick view of the overall plausibility of the synthetic data but may also provide insight for selecting a suitable quantitative assessment metric for subsequent evaluation. Visualizations have not been termed as a metric for realism evaluation in the literature, but we consider them important tools that provide a quick view of the SD quality and can offer leads for more detailed assessment. The majority of the studies included in this review have used some form of visual assessment of realism as depicted in Table 9.

5.3.2. Privacy preservation

The second quality attribute of synthetic data is privacy preservation. Synthetic data alleviates privacy concerns of published

Table 8
Utility metrics for SDG realism employed by included studies.

Utility metric	Articles employing this metric
Performance Agreement[a]	Classification [21,23,28,38–40,42,78,84,93,94,97,97,99,101,103,105,107,110,133,148,150] Regression [31,42,93] Clustering-based Classification Accuracy [29]
Data Augmentation Test	[39,41–43,92,148]
Model Congruence	Model Parameter Agreement [74,76,77,99] Feature Ranking Agreement [31,40,97,99,110] Synthetic Rank Agreement [97] Response Agreement [83,108]
Field Experiments	Clinical Trial Selection [68] Testing of Clinical Support System [36] Research Studies Replication [28] Teaching and Training [28] Clinical Text De-Identification [148]

[a]Performance Agreement includes different settings as TSTR, TRTS or the classical TRTR vs. TSTS comparison

Table 9
Qualitative realism assessment methods used in included studies.

Metric		Studies employing the method
Expert Evaluation		[24,35,36,39,42,70,71,81,99,103,144,148,150]
Visual Plausibility	Basic visualizations	Marginal Distributions Histograms [76] Distribution Histograms [42,77,80,82,85,89,95,100,102,147,149] Mean Vectors [101] Feature Pairs Distribution graphs [42] Cross-Type Conditional Distribution Plot [91] Pairwise Correlation heatmaps [39,40,80,99,101,143,146,147] Spearman correlation [42,85]
	Condensed projections	NMDS [42,143] PCA Plots [28,38,93,146] t-Sne [103] KDE [27] UMAP [93]

data by synthesizing values. Following this notion, several earlier studies assumed inherent privacy guarantees of SD (Table 10) and as a result, many authors skipped a rigorous privacy evaluation of their methods. Fig. 12 shows a high-level classification of the privacy preservation methods employed by included studies. Although privacy may not be a core consideration for synthetic data in some application domains such as test data, it is an essential concern for synthetic medical data [6]. Fully synthetic datasets contain artificially created records and thus do not have a one-to-one mapping with a real record making identity disclosures meaningless [155]. However, Emal et al. [56] assert that the inclusion of rare instances in SD can potentially reveal the identities of individuals with similar records in the real data and term this as "meaningful identity disclosure" risk. The authors carry out an empirical evaluation of meaningful identity disclosure risks with two synthetic datasets and demonstrate that this risk is significantly below the accepted threshold but urge further investigations. Nevertheless, synthetic instances that are too similar to the real records can be exploited to make inferences about the original data used to train the generative model [28]. Knowledge-driven synthetic data do not use any private information for the generative modeling; hence, privacy assessment is not relevant to such approaches. However, the DD and HBD synthetic data are susceptible to disclosure attacks [156]. Synthetic datasets are challenged by following privacy threats:

- **Membership Inference (MI) or Presence Attacks:** Membership or presence disclosure happens if an adversary can conclude with certain confidence that an individual's record was used to train the generative model. Membership disclosure can reveal some aggregate-level information about individuals. For example, if Bob has sufficient background knowledge about Alice and by looking at the synthetic records for skin-cancer patients, he is assured that Alice's record was in the training set, then he learns with some confidence that Alice also has that disease. Membership disclosures will eventually lead to attribute disclosures and thus recent studies emphasize accurate risk assessment of presence disclosures before releasing synthetic data [157]. Membership disclosure assessment is carried out through simulation of such attacks by finding synthetic records that match with real instances which are then tagged as members of the real training set. A high proportion of true positives, in this case, indicates a higher risk of membership disclosure. The common scheme for such simulation involves an attack dataset that includes real instances from training and holdout data in equal proportion [155,158]. However, Eman et al. [157] highlighted a discrepancy in this method and demonstrated that the proportion of real records from the training partition should match with the sampling fraction of the real training data from the population for accurate estimation. Zhang et al. [159] established that partially synthetic data are more susceptible to membership inference as compared to fully synthetic data which are fairly resistant to these attacks.
- **Attribute Disclosure (AD) Attacks:** Attribute disclosure happens with overfitted models when the adversary can get narrow bounds on the real values of the sensitive data. Researchers take attribute disclosure as a high-priority privacy consideration in comparison with membership disclosure as the adversary only needs to know a subset of original attributes to infer the remaining values from similar records in the synthetic data [156].

Table 10
Privacy metrics employed by the DD and HBD methods.

Metric classification	Privacy metric	Studies employing the metric	
Integrated	Wrapper	MI	
		Differential Privacy [75–77,84,97,99,100,102,103] Renyi DP [104]	
	Embedded	AD	MI
		v-dispersion [73]	Moment Accountant DP [98,108] $\in$-Identifiability [101] Synthetic l-diversity [76,77] probabilistic K-anonymity [74]
PostHoc	Nearest Neighbors	AD	MI
	(NN) Metric	NN attribute estimation [22,81,91,105] Reproduction Rate [21,24,42,79,90,150] ROUGE-N [148]	NN Similarity [80,83,88,90,91,100,146,147] (MI) AA Privacy Loss [28,38,93,133] Membership Inference by attribute estimation [133] Outlier Similarity [21,23,143,147]
	Distributional Separation		Perplexity Distributions Similarity [105] Mean Discrepancy [40,111]
Assumed	Inherited Default Built-In (KD)	[39,41,43,82,85–87,89,94,95,106,107,149] [23,27,29,31,35,49,78,79,92,110,144,145] [25,32,36,68–72]	

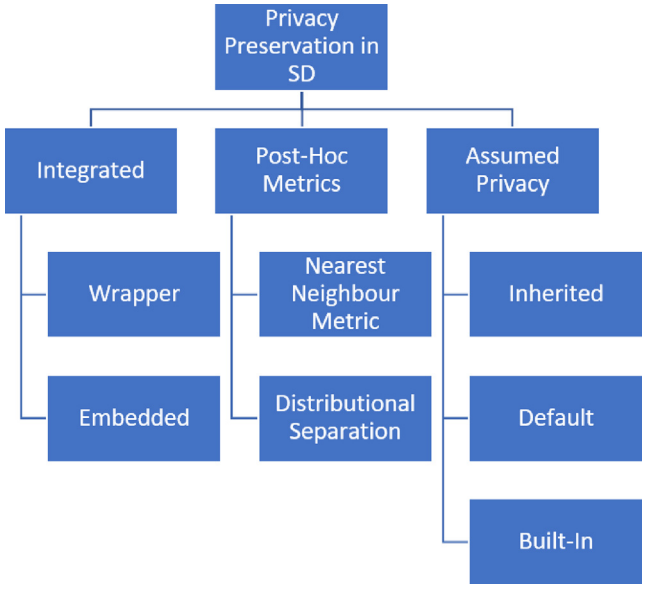

Fig. 12. High-level classification of SD privacy preservation methods.

A careful privacy evaluation of synthetic data is vital to thwart any disclosure risks. Membership disclosures can lead to inferences about attribute values and vice versa. Based on this, most of the privacy metrics, defined in the literature, provide a combined privacy assessment against these threats. In the following section, we present a discussion of the various privacy metrics used for the synthetic dataset. Table 10 enlists the SD privacy metrics along with the studies that have employed them.

A. Privacy Metrics

In this section, we discuss the different privacy metrics proposed in the literature for the privacy assessment of synthetic data. Fully synthetic data do not face the identity disclosure risk as no direct link between synthetic and real records exists, but synthetic records that are too similar to the original ones open doors for membership and attribute disclosures [28]. We noted that although researchers have employed diverse methods for privacy evaluation of the synthetic data, all of them can be grouped into two main classes. (1) Privacy is integrated into the generative process and the privacy mechanism controls, as well as measures the disclosure risk (2) Post synthetic privacy evaluation, is conducted to quantify the disclosure risk. In addition to these, there was a third group of studies, for which authors assumed privacy as being inherited by the generated model. We call this third category "Assumed Privacy". Fig. 12 depicts the high-level classification of SD privacy preservation.

1. Integrated Privacy: The metrics in this category are integrated into the SDG process and regulate the privacy mechanism at the time of data synthesis. We further classify integrated privacy as ***Wrapper*** or ***Embedded*** based upon the mode of operation. In the wrapper mode, privacy protection is added as a separate mechanism running in parallel with the data synthesis which performs some transformations over the generation at certain points in time. An example of such a mechanism is noise injection or gradient clipping of the generator. The embedded privacy, on the other hand, includes privacy interventions that are built into the objective function of the GM. Privacy mechanisms can be adapted to work in any of these modes with some updates to the generative process. **Differential Privacy *(DP)*** is a de-facto privacy principle and has also been used frequently for privacy-preserving SDG, mostly as a Wrapper (Table 10). Some studies implemented DP in embedded mode using the *Moments Accountant* strategy [98,108]. $\in$-*Identifiability* [101] embeds an identifiability loss term into the GM to control the amount of information about the original distribution being injected into the generator. The notion of identifiability in SD is inspired by the popular anonymization metric k-anonymity to alleviate membership disclosure risks. Some other anonymization metrics such as l-diversity, have also been adapted for SD as shown in Table 10. The anonymization-inspired metrics are generally implemented in embedded mode.
2. Post-Hoc Privacy Metrics: The metrics in this category are based on post-processing of the generated instances to measure disclosure risks. Various forms of ***Nearest Neighbors Metrics*** have been employed in different studies (Table 10) which measure proximity or similarity between synthetic and original records as a basis to determine privacy preservation. However, what proximity or similarity means is subjective to the target scenario. Different distance metrics e.g., Hellinger, Euclidian, and Hamming,

to name a few, have been used in the literature for this purpose. *Adversarial Accuracy Privacy Loss* was introduced in [28,38] to collectively measure utility and privacy based on the nearest neighbor metric. Researchers argue that synthetic records that are too similar to the original ones leak substantial information about the real data [90,101]. The worst case is the *duplication* of a real record in SD which is essentially an indicator of memorization induced in the GM because of overfitting [118]. Depending upon the severity of overfitting, models may reproduce complete or partial records *Reproduction Rate* is the proportion of duplicated records in the synthetic dataset and has been used as a measure of disclosure risk [90]. Outliers in the data can reveal significantly more information than regular records and may potentially cause meaningful identity disclosures under certain scenarios [56]. *Outlier Similarity metric* assesses the risk of disclosures from synthetic outliers that are close to their real counterparts. Attackers with some background knowledge about a patient might exploit the nearest neighbors in the synthetic data to *estimate the values for unknown attributes*. Yale et al. [133] extended the attribute estimation to build a membership inference attack model [133]. The variants of NN metric operate at the record level and formulate an aggregate privacy loss/preservation score. Macro measures of privacy preservation through distributional similarity between SD and training and hold-out datasets have also been used in the literature [40,105,111]. If the distributions of SD and hold-out data are closer than those of SD and training data, this indicates lesser memorization and thus lesser disclosure risks [105].

3. Assumed Privacy: We noted that, in addition to the above metrics, some researchers assumed privacy guarantees for their SD under two scenarios. (1) They enhanced an existing generative model for which privacy guarantees were already established in an earlier study; (2) They argued that certain machine learning models (such as. RDT [31]), have built-in randomization mechanisms to safe-guards against privacy leaks thus minimizing disclosure risks from generated data. We collectively call these assuming *Inherited Privacy* (Table 10). In the case of KD approaches, privacy is built-in into the SD as no sensitive information is ever used. Surprisingly, some studies employing DD or HDB approaches also assumed built-in privacy arguing that privacy protection is default to syntheticity. We group such studies under a separate category called the *default* privacy category (Table 10, Fig. 12). Close to 50% of the included studies assumed privacy including the KD approaches with built-in privacy as shown in Fig. A.2, Appendix A. However, it is important to note that more recent literature on SDG deems it mandatory to perform an explicit privacy assessment of synthetic data before sharing [156]. We performed a closer analysis of the DD approaches under the assumed category in terms of the GM employed and model it as a sunburst chart in Fig. 13. It is clear from Fig. 13, that inherited privacy is mostly assumed with ML-based and particularly GAN-based GMs, while a higher proportion of classical methods resorted to the default privacy assumption.

6. Discussion and synthesis

In this section, we highlight the novelty and contribution of this review followed by a discussion of our findings. The contribution of this study is three-fold. First, we have presented a narrative review of 70 peer-reviewed research articles discussing synthetic medical data generation for privacy-safe data publishing and highlight the methods and evaluation techniques used for SD. Secondly, we have provided a classification of existing SDG approaches for medical data generation. Thirdly, a classification of the different SD quality evaluation metrics is also furnished.

The increasing number of studies each year (Fig. A.1 – Appendix A) is indicative of the recognition of SD as a prospective substitute for real data. However, the adoption of SD in practical applications is progressing at a slower pace and few real-life use cases of medical SD (e.g., classroom training [28], designing of diabetes foot insole [27]) were found in the included studies. Nevertheless, several synthetic datasets have been used for various machine-learning tasks (Table B.5 – Appendix B) with promising results. Although the findings from these laboratory experiments with SD may not be directly applicable to the clinical practice in their present form, they do pave the way for this eventual goal. In light of our findings, we propose a way forward in medical SDG research on four fronts.

6.1. Hybrid methods — the way forward

Medical data is highly dimensional with a complex inter-feature correlation which presents a big challenge for the generative modeling both in KD and DD modes. KD based approaches face incompleteness and omission issues because manual model curation and DD methods require large volumes of high-quality real data to learn from. The hybrid approach presents an intuitive choice for synthetic health data generation because of the flexibility and control of the KD automation and the efficiency of DD elements. Existing HBD methods have dominant KD components which demand extensive manual intervention. New hybrid approaches with an optimal mix of data and knowledge are required to exploit the best of both methods. Methods similar to the "advise" mechanism [49] to generative models are a promising way forward.

6.2. Machine-friendly representation of domain knowledge

The importance of domain knowledge in constructing and evaluating synthetic health records cannot be undermined. However, for the most part, this is a manual process that is both inefficient and error-prone. Many phases of this labor-intensive process can be automated by providing a homogeneous machine-ready representation of domain knowledge consolidated from different sources. There have been many such efforts in the literate e.g. in form of disease progression models [160]. Integrating such models with the hybrid synthetic data generation mechanism can improve the overall quality of SD.

6.3. Granularity over spectrum

EHRs provide the richest source of information for medical research and discovery but are unavailable to the research community for privacy concerns. High-quality synthetic EHRs can unleash the huge potential for medical research, but huge volumes of real data are needed to accomplish this. The diversity of information in EHRs is among the top agents challenging the realism of synthetic data. Multimorbidity is an important and common concept in healthcare [161], which is a source of complexity in patient EHRs. Success with the conditional synthesis of EHR [90] is a prospective direction of research that can reduce the overall complexity of the problem by targeting a specific medical specialty. Conditional synthesis can work with smaller training datasets while still leveraging sufficient coverage. Targeting a wider spectrum of health conditions in the synthetic datasets ends up compromising the granularity or detail of information, more specifically, the temporal and sequential attributes. The finer granularity of synthetic data is a more important concept,

Fig. 13. Classification of studies with assumed privacy.

in terms or utility than the spectrum of health conditions represented in it. Aggregated data are suitable for certain machine learning tasks e.g. disease prediction [81], yet, more detail is needed for clinical analysis and advanced medical research [120]. Advancing research in this direction will facilitate future research targeting multimorbidity. Research is needed to produce quality synthetic longitudinal records for specific phenotypes as a starting step to progress towards full-blown EHRs in the future.

6.4. Privacy evaluation

We noted that privacy has been given lesser attention compared to realism in the evaluation of SD quality, under the assumption that SD is inherently resistant to disclosure. This is evident from Fig. 13 which shows a large proportion of studies lacking rigorous privacy evaluation. However, literature has established that SD is not completely impervious to privacy attacks and mandated careful privacy assessment of SD before release [162]. It should also be noted that privacy cannot be viewed in isolation from utility due to the inverse association between the two [163]. This demands a tradeoff between privacy and utility for optimal quality of SD for a given application. More research is required to formalize the relationship between these quality attributes.

6.5. Synthetic data quality evaluation framework

The growing volume of literature on medical SDG also brings with it an array of new terms, concepts, techniques, models, and metrics. Different studies have used various quality evaluation metrics and derived inferences from the results. However, the lack of standardized representation of quality metrics hinders a fair and conclusive comparison between the different methods and thus selection of a suitable method remains a challenge. There is a need to develop a quality evaluation framework for synthetic data which will serve two purposes.

1. Provide a platform for practitioners to assess the utility of synthetic data for their target applications and select a suitable method without worrying about technical intricacies. This will facilitate the wider adoption of SD in practice.
2. Establish benchmarks and provide guidelines for new research in this domain and for contributing towards the maturity of the field.

We also mention some recent advances in SDG evaluation framework including the work of Eman et al. [162] to study the effectiveness of multivariate statistical metrics in ranking an SDG method, by comparing it against the task-specific utility assessment. The authors performed extensive experimentation using six common statistical metrics over 30 datasets and concluded that Hellinger distance provides the closest estimates of SD utility for classification tasks. Similarly, Yale et al. [163] put forth a selection framework to rank SDG models for different use cases. The study investigated popular GAN models for SDG using two public healthcare datasets and concluded that no specific method is unequivocally superior in all criteria for utility and privacy, yet some models will work better for certain use cases. The framework

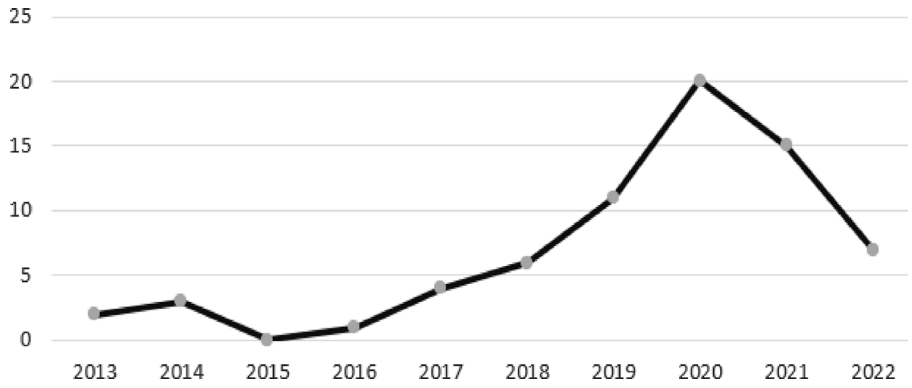

Fig. A.1. Year-wise count of included studies (automated search till Jul 30, 2022).

operates in three phases: (1) Synthetic Data generation through selected methods (2) Metric-specific ranking lists against each metric and (3) Use-cases specific model rankings. The final output of this process is the model recommendation for each use case.

7. Limitations

Several limitations should be noted. The search results only include peer-reviewed journals and conferences, and relevant work may have been missed out from non-peer review sources. With synthetic data generation being a nascent area of research, much of the research is published in repositories. However, forward snowballing has been employed to include recent work taking into consideration the reliability of the primary studies which may be absent in non-peer-reviewed sources. The data extracted is done exclusively by the first author of this study and may be subject to human biases. However, the iterative process of extracting and discussing with peer researchers may allay human biases. Also, the data extracted from each primary study has been presented in the repository which makes it possible to scrutinize the researchers' judgment and ensure the repeatability of the review results. Another limitation is inherent in the choice of the literature review in this study. We have adopted a narrative review in this study which has highlighted different synthetic data generation methods in health care along with their strengths and weaknesses. And we have presented an overall understanding of the phenomenon in the above-mentioned domain which is the main motive of this type of review. However, it falls short of critically analyzing the literature and fails to challenge problematic issues that may be present in the literature.

8. Conclusion and future directions

This paper presented a narrative review of the state of the art in synthetic medical data generation to provide a quick reference guide for the researchers working in this area. We reviewed 70 peer-reviewed articles that passed our selection criteria. Synthetic data is rapidly gaining attention as a privacy-safe data-sharing method that can unravel the potential for research and innovation in medical informatics. Synthetic data with high realism can substitute real datasets in many applications including algorithm testing and validation, technology evaluation, teaching and training, and data science competitions to name just a few, without compromising the privacy of individuals. The knowledge-driven synthetic data generation methods provide high privacy guarantees with the flexibility to fine-tune the synthetic sample characteristics but are hindered by sheer dependence on the manual specification of the generative model. Machine process-able representation of domain knowledge (disease progression models, clinical procedure guidelines) can enable automated integration and unification of domain information from heterogeneous sources into the generative models to improve the overall efficiency and efficacy of knowledge-driven synthetic data generation. Data-driven approaches suffer from the 'boot-strap paradox' [118] and face serious challenges in learning an appropriate generative model for high-dimensional and highly inter-correlated medical data. Hybrid approaches can alleviate the task by mitigating the data-related discrepancies through abundantly available domain knowledge. Researchers have published promising results of experiments with the synthetic generation of parts of longitudinal EHRs. Still, generating comprehensive longitudinal records with co-morbidities remains an open challenge. Hybrid approaches provide the necessary framework for longitudinal EHRs with a flexible balance between realism and privacy, but they have not been studied sufficiently. The existing realism metrics evaluate the univariate, multivariate, and utility-based realism as independent operations which can potentially give a misleading view of the true situation. Comprehensive metrics combining synchronized evaluations of structural and utility realism are required. Moreover, embedded assessment of realism for data-driven generative processes has the potential to produce more accurate estimates.

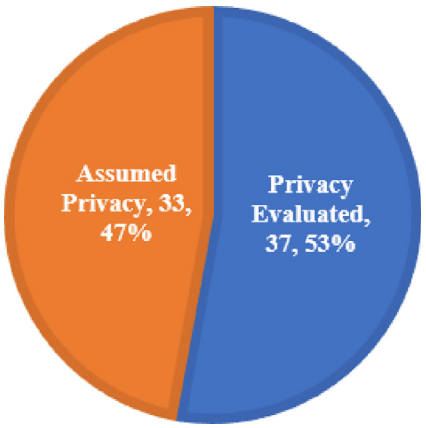

Fig. A.2. Privacy assessment in included studies — assumed vs. evaluated proportion.

Table B.1
Classification of included studies according to source of ground truth.

SDG approach	Studies in the category
Data Driven Methods	[22,23,28,31,38–43,73–90,92–95,97–108,110,111,133,149,150]
Knowledge Driven Methods	[25,32,36,68–72]
Hybrid Methods	[21,24,27,29,35,49,143–148]

Table B.2
Medical datasets utilized by included studies and their granularity and availability. Restricted datasets have public access after fulfilling some regulatory requirements.

Dataset	Access	Dataset granularity (Original)	Studies utilizing this dataset
UCI — Contraceptive Method Choice Data [112]	Public	Snapshot	[108]
UCI — Thyroid [112]	Public	Snapshot	[73,74]
UCI — Diabetes [112]	Public	Snapshot	[27,29,31,73,80,108]
UCI — Mammographic Mass [112]	Public	Snapshot	[108]
UCI — Breast Cancer [112]	Public	Snapshot	[31,84,108]
UCI — Heart Disease Data Set [112]	Public	Snapshot	[80]
UCI — Indian liver patient dataset [112]	Public	Snapshot	[42]
UNOS — Heart wait-list [164]	Public	Snapshot	[97,101]
UNOS — Heart Transplant [164]	Public	Snapshot	[101]
UNOS — Lung Transplant [164]	Public	Snapshot	[101]
BSA Inpatient Claims [165]	Public	Snapshot	[76]
National Long Term Care Survey (NLTC)	Public	Snapshot	[75]
CDC NSSP BioSense Dataset [166]	Restricted	Longitudinal	[24]
CDC NCHS Heath Dataset (selected features/)	Public	Snapshot	[83]
Project Data Sphere (PDS) Clinical Trials [167]	Public	Snapshot	[110]
SEER Data set [113]	Public	Snapshot	[22]
PTB Diagnostic ECG [168]	Restricted	Aggregate	[104]
Philips eICU Database	Restricted	Timeseries/Longitudinal	[100]
MAGGIC Dataset [169]	Public	Aggregate/Snapshot	[101]
Nephrotic Syndrome Dataset	Private	Aggregate/Snapshot	[49,84]
Kaggle cervical cancer dataset [112,170]	Public	Aggregate	[84,97,104]
Kaggle cardiovascular dataset	Public	Aggregate/Snapshot	[84,104]
Kaggle Medical Insurance Cost Dataset	Public	Snapshot	[93]
PIMA Indian Diabetes Dataset [171]	Public	Snapshot	[93]
UCI — Parkinson's Telemonitoring [112,172]	Public	Timeseries	[31]
UCI — Epileptic Seizure Recognition [112,173]	Public	Timeseries	[84,88,97,104]
UCI EEG Eye State [112]	Public	Time Series	[107]
NHIRD Public Release, Taiwan [174]	Public	Timeseries/Snapshot	[86,87]
MIT-BIH Arrhythmia [175,176]	Public	Time Series	[89,104,107]
BIDMC PPG and Respiratory [176,177]	Public	Time Series	[89,107]
Siena Scalp EEG Database [178,179]	Public	Time Series	[89]
MIMIC- II, III [116,180]	Restricted	Longitudinal	[28,35,39,49,81,82,84–88,94,98,99,103,104,133,145,147]
Vanderbilt University Medical Center Synthetic Derivative (VUMC- SD) [181]	Restricted	Longitudinal	[90,91,105,106]
Patient Discharge Data, California [182]	Public	Longitudinal	[74,77]
Private datasets from hospitals (Not Published)	Private	Longitudinal	[38,43,92,95,149,150]
Privately collected dataset from individuals (Not published)	Private	Timeseries	[102]
Private EHR text	Private	Text	[148]
Clinical Practice Research Datalink (CPRD) [183]	Restricted	Longitudinal	[21,42,143,147]
CERNER Health Facts [184]	Public	Longitudinal	[40]
Gene Expression Data Set DNA microarray data [41]	Public	Snapshot	[41]
Treadmill maximal exercise test (TMET) [185,186]	Restricted	Timeseries	[146]
Canadian Covid −19 case dataset	Restricted	Longitudinal	[111]
St. Louis Children's Hospital (SLCH), Virtual PICU Systems (VPS, LLC) Pediatric Intensive Care Unit (PICU) registry [187]	Restricted	Timeseries	[23]
United States National COVID Cohort Collaborative (N3C) [50]	Restricted	Aggregate/Snapshot	[78,79]
National Institute of Health (NIH) All of Us Research Program.	Restricted	Longitudinal	[106]

Table B.3
Classification of included studies according to the granularity of synthetic data.

Synthetic data set granularity	Studies
Snapshot	[21,22,28,31,36,38,41,42,68,73–80,83,84,93,97,101,104,108,110,111,133,143]
Aggregate	[23,25,40,43,49,81,82,85–88,90,91,98,102,104]
Longitudinal	[24,32,69–72,103,105,106,144,145,150]
Timeseries	[27–29,38,39,70,89,94,95,99,100,103,107,133,146,147,149]
Text	[35,92,148]

Table B.4
Classification of included studies according to feature types.

Data types	Studies
Discrete only	[22,35,43,75–77,81,82,86,87,90,92,98,103,105,106,148,150]
Continuous only	[39,73,89,93,95,99,107,149]
Mixed	[21,23–25,27–29,31,32,36,38,40–42,49,68–72,74,78–80,83,85,88,91,94,97,100–102,104,108,110,111,133,143–147]

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Appendix A

See Figs. A.1 and A.2.

Appendix B

See Tables B.1–B.6.

Table B.5
Synthetic dataset generated by included studies and corresponding use cases.

Studies	Synthetic datasets	Uses cases as in the study
[68]	Breast Cancer Patients with first diagnosis	Case selection for clinical trial
[36]	Patient Data for Depression	Integration with Clinical support systems Testing of Clinical Support System Privacy Preserving Data Sharing
[73]	Thyroid patient dataset Diabetes patient dataset	Privacy Preserving Data Sharing
[74]	Thyroid patient dataset Neoplasm dataset	Privacy Preserving Data Sharing
[71] [69,70] [25]	Longitudinal Records for Type II Diabetes Longitudinal Records for Delivery Episodes Dataset Type 2 Diabetes Mellitus Dataset	Privacy Preserving Data Sharing for non-clinical use Privacy Preserving Data Sharing for non-clinical use Urban Planning and Support
[72]	Chronic Obstructive Pulmonary Disease (COPD) Dataset Colorectal Cancer Dataset Hip/Knee Replacement Complications Dataset High Blood Pressure Patients Dataset	Privacy Preserving Data Sharing for non-clinical use
[75]	Quality of care dataset	Privacy Preserving Data Sharing for non-clinical use
[42]	Liver disease data set Cardiovascular disease data	Liver disease Prediction 10-years CVD risk prediction Dataset Real Dataset Augmentation/Balancing
[76] [80] [77] [83]	Patient Discharge Data Aggregated Patient Diagnosis Dataset Medicare Claims Data Health Data set	Predicting Hospital Stay Cost Privacy-Preserving Data Sharing Predicting Hospital Stay Length Privacy Preserving data sharing
[188]	Heart Disease Data Diabetes patient data Aggregated Patient diagnosis History	Privacy Preserving data Sharing
[81] [82] [85]	Aggregated Patient diagnosis History Aggregated Patient diagnosis History with Demographics Aggregated Patient Diagnosis History	Privacy Preserving Data Sharing for Research Privacy Preserving Data Sharing for Research Privacy Preserving Data Sharing for Research
[86,87]	Aggregated Patient diagnosis History Aggregated Patient diagnosis History and procedures	Privacy Preserving Data Sharing for Research
[22]	Breast Cancer Data Respiratory Cancer Data, Leukemia dataset	Privacy Preserving Data Sharing for Research
[43]	Pediatric Data set	Treatment Recommendation
[39]	Sepsis medication dosage Dataset Diabetes medication dosage Dataset	Medication Dosage Recommendation
[88]	Aggregated Patient Diagnosis Epileptic Seizure Recognition	Recognizing Epileptic activity
[89] [90]	Biomedical signal Data Aggregate Patient Diagnosis Dataset	Privacy Preserving Data Sharing Privacy Preserving Data Sharing
[93]	Medical Cost Dataset Diabetes Dataset	Disease Prediction (Diabetes) Insurance Cost Prediction
[40]	Aggregated Patient Diagnosis History with lab values	Data Augmentation Disease Prediction
[41]	Gene Expression data for Colon and Breast Cancer tissues	Data Augmentation
[28,38,95,149]	Aggregated ICU Patient Data (demographics, disease, vitals, procedures) Autism Spectrum Disorder	48 h Mortality Prediction Academic Use Replication of Research Studies
[133]	Aggregated ICU Patient Data (demographics, disease, vitals, procedures)	48 h Mortality Prediction
[94]	Aggregated Patient Diagnosis with Time Series data	In-Hospital Mortality Prediction Decompensation Prediction Disease Classification
[97]	Epileptic Seizure Dataset Chronic Heart Failure Dataset Organ Transplant Dataset Cervical Cancer Dataset	Disease/Outcome Prediction
[98]	Aggregated Patient Diagnosis History Dataset	Privacy-Preserving Data Sharing
[99]	Heart-failure dataset SPRINT Data set	Predicting heart failure Predicting treatment arm
[100]	EEG Dataset	Data Augmentation Privacy-Preserving data sharing

(*continued on next page*)

Table B.5 (*continued*).

Studies	Synthetic datasets	Uses cases as in the study
[101]	Organ Transplant Dataset Chronic Heart Failure Dataset	Predicting 3-year mortality
[103,150]	Longitudinal Patient Data	Privacy-Preserving Data Sharing Disease Prediction
[105] [106] [107]	Longitudinal Patient Data for selected diseases Longitudinal Patient Data with disease and procedure codes ECG, PPG, EEG, Balleostrocardiogram signals	Diagnosis Forecast Privacy-Preserving Data Sharing Privacy-Preserving Data Sharing
[108]	Breast Cancer Dataset Diabetes Dataset	Breast Cancer Detection Diabetes Detection
[31]	Breast Cancer Dataset Parkinson's Dataset Diabetes Dataset	Privacy-Preserving Data Sharing
[21] [110] [144] [143]	Primary Care Dataset Clinical Trial Datasets Longitudinal Patient Records Cardiovascular Disease dataset	Stoke Prediction Clinical Trial Selection Privacy-Preserving Data Sharing Privacy-Preserving Data Sharing
[24]	Longitudinal Patient Data Tularemia outbreak dataset	Detecting an epidemic outbreak
[145] [27,29] [147]	Patient Records with Timeseries data Diabetic Foot Data Blood Pressure Time Series Data	Privacy-Preserving Data Sharing Designing Diabetic Foot Treatment insole Privacy-Preserving Data Sharing
[49]	Nephrotic Syndrome Dataset Aggregated Patient Diagnosis Dataset	Predicting Nephrotic Syndrome Developing Clinical Support System Privacy-Preserving Data Sharing
[35]	Clinical Notes for Psychiatric Patients	Privacy-Preserving Data Sharing
[148]	Clinical Notes	Privacy-Preserving Data Sharing Training Models for automatic De-identification of PHIs
[102] [146]	Fitbit smart health dataset Treadmill maximal exercise test dataset	Privacy-Preserving Data Sharing Privacy-Preserving Data Sharing
[111]	Covid-19 case dataset [189]	Privacy-Preserving Data Sharing Data Analysis
[104]	Aggregated Patient Diagnosis Dataset Disease Specific Datasets (cardiovascular, cervical cancer, arrhythmia) EEG dataset	Privacy-Preserving Data Sharing
[92]	Medical Text	Privacy - Preserving Data Sharing
[23]	Pediatric ICU Dataset Sepsis Dataset Chlamydia Dataset	Privacy - Preserving Data Sharing Sepsis Prediction Data Analysis for Public Health
[91]	Aggregated Patient Diagnosis and Procedures Dataset	Privacy - Preserving Data Sharing
[78,79]	COVID Geo-spatial Dataset	Privacy – Preserving Data Sharing 14-day admission risk prediction Geo Spatial Data Analysis for Public Health and epidemiology

Table B.6
Classification of studies based on the Generative Model employed.

Model category	Model names and studies employing them				
Structural 1	XML Descriptor	Decision tables			
	[36,68]	[144]			
Behavioral 3	[24,25,27,29,32,35,69–72,145,146]				
ML based models	GANs 3	Auto-encoder	Neural Networks	Decision Trees/Randomized DTs 2	
	[22,28,38–41,43,49,81–95,97–106,133,149]	[103,108]	[35,107,148]	[31,110,111]	
Classical models	Mixture models	KDE 3	Copula	Monte Carlo	Bayesian Networks
	[73,74]	[23,78,79]	[42]	[76,77]	[21,42,75,80,143,147]

References

[1] R. Gururajan, A. Hafeez-Baig, An empirical study to determine factors that motivate and limit the implementation of ICT in healthcare environments, BMC Med. Inform. Decis. Mak. 14 (1) (2014) 98, http://dx.doi.org/10.1186/1472-6947-14-98.

[2] Faisal Said Al Habsi, Dr. Elango Rengasamy, Managing obsolescence and prolonging the useful life of desktop computers – an exploratory analysis, Int. J. Manage. 11 (6) (2020) 293–322.

[3] M.R. Cowie, et al., Electronic health records to facilitate clinical research, Clin. Res. Cardiol. 106 (1) (2017) 1–9, http://dx.doi.org/10.1007/s00392-016-1025-6.

[4] W.G. van Panhuis, et al., A systematic review of barriers to data sharing in public health, BMC Public Health 14 (1) (2014) 1144, http://dx.doi.org/10.1186/1471-2458-14-1144.

[5] E.S. Dove, M. Phillips, Privacy law, data sharing policies, and medical data: A comparative perspective, in: A. Gkoulalas-Divanis, G. Loukides (Eds.), Medical Data Privacy Handbook, Springer International Publishing, Cham, 2015, pp. 639–678, http://dx.doi.org/10.1007/978-3-319-23633-9_24.

[6] B. Malin, K. Goodman, Between access and privacy: Challenges in sharing health data, Yearb Med. Inform. 27 (1) (2018) 55–59, http://dx.doi.org/10.1055/s-0038-1641216.

[7] F. Li, X. Zou, P. Liu, J.Y. Chen, New threats to health data privacy, BMC Bioinformatics 12 (12) (2011) S7, http://dx.doi.org/10.1186/1471-2105-12-S12-S7.

[8] C. for, O. Rights (OCR), HIPAA for Professionals, HHS.gov, 2015, https://www.hhs.gov/hipaa/for-professionals/index.html (accessed Oct. 25, 2020).

[9] General data protection regulation (GDPR) – official legal text, general data protection regulation (GDPR), 2021, https://gdpr-info.eu/ (accessed May 20, 2021).

[10] M. Jayabalan, M.E. Rana, Anonymizing healthcare records: A study of privacy preserving data publishing techniques, Adv. Sci. Lett. 24 (3) (2018) 1694–1697, http://dx.doi.org/10.1166/asl.2018.11139.

[11] A. Pawar, S. Ahirrao, P.P. Churi, Anonymization techniques for protecting privacy: A survey, in: 2018 IEEE Punecon, 2018, pp. 1–6, http://dx.doi.org/10.1109/PUNECON.2018.8745425.

[12] S.M. Bellovin, Privacy and synthetic datasets, 39.

[13] F. Skopik, G. Settanni, R. Fiedler, I. Friedberg, Semi-synthetic data set generation for security software evaluation, in: 2014 Twelfth Annual International Conference on Privacy, Security and Trust, 2014, pp. 156–163, http://dx.doi.org/10.1109/PST.2014.6890935.

[14] S. Popić, I. Velikic, N. Teslic, B. Pavkovic, Data generators: a short survey of techniques and use cases with focus on testing, 2019, http://dx.doi.org/10.1109/ICCE-Berlin47944.2019.8966202.

[15] Synthetic datasets for statistical disclosure control - theory and implementation | Jörg drechsler | Springer, 2020, Accessed: Oct. 18, 2020. [Online]. Available: https://www.springer.com/gp/book/9781461403258.

[16] C. Lee, et al., Big healthcare data analytics: Challenges and applications, in: S.U. Khan, A.Y. Zomaya, A. Abbas (Eds.), Handbook of Large-Scale Distributed Computing in Smart Healthcare, Springer International Publishing, Cham, 2017, pp. 11–41, http://dx.doi.org/10.1007/978-3-319-58280-1_2.

[17] R.J. Chen, M.Y. Lu, T.Y. Chen, D.F.K. Williamson, F. Mahmood, Synthetic data in machine learning for medicine and healthcare, Nat. Biomed. Eng. 5 (6) (2021) 493–497, http://dx.doi.org/10.1038/s41551-021-00751-8.

[18] E. Borycki, Trends in health information technology safety: From technology-induced errors to current approaches for ensuring technology safety, Healthc. Inform. Res. 19 (2) (2013) 69, http://dx.doi.org/10.4258/hir.2013.19.2.69.

[19] O. Vovk, G. Piho, P. Ross, Anonymization methods of structured health care data: A literature review, in: Model and Data Engineering, Cham, 2021, pp. 175–189, http://dx.doi.org/10.1007/978-3-030-78428-7_14.

[20] S. James, C. Harbron, J. Branson, M. Sundler, Synthetic data use: exploring use cases to optimise data utility, Discov. Artif. Intell. 1 (1) (2021) 15, http://dx.doi.org/10.1007/s44163-021-00016-y.

[21] A. Tucker, Z. Wang, Y. Rotalinti, P. Myles, Generating high-fidelity synthetic patient data for assessing machine learning healthcare software, Npj Digit. Med. 3 (1) (2020) 1, http://dx.doi.org/10.1038/s41746-020-00353-9.

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

[23] R.E. Foraker, et al., Spot the difference: comparing results of analyses from real patient data and synthetic derivatives, JAMIA Open 3 (4) (2021) 557–566, http://dx.doi.org/10.1093/jamiaopen/ooaa060.

[24] A.L. Buczak, S. Babin, L. Moniz, Data-driven approach for creating synthetic electronic medical records, BMC Med. Inform. Decis. Mak. 10 (1) (2010) 59, http://dx.doi.org/10.1186/1472-6947-10-59.

[25] Y. Liu, R. Stouffs, Y.L. Theng, Development of synthetic patient data to support urban planning for public health, in: Presented At the ECAADe 2020: Anthropologic : Architecture and Fabrication in the Cognitive Age, Berlin, Germany, 2020, pp. 315–322, http://dx.doi.org/10.52842/conf.ecaade.2020.1.315.

[26] A.H. Pollack, T.D. Simon, J. Snyder, W. Pratt, Creating synthetic patient data to support the design and evaluation of novel health information technology, J. Biomed. Inform. 95 (2019) 103201, http://dx.doi.org/10.1016/j.jbi.2019.103201.

[27] J. Hyun, S.H. Lee, H.M. Son, J.-U. Park, T.-M. Chung, A synthetic data generation model for diabetic foot treatment, in: Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications, Singapore, 2020, pp. 249–264, http://dx.doi.org/10.1007/978-981-33-4370-2_18.

[28] 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.

[29] J. Hyun, et al., Synthetic data generation system for AI-based diabetic foot diagnosis, SN Comput. Sci. 2 (5) (2021) 345, http://dx.doi.org/10.1007/s42979-021-00667-9.

[30] S. Gerke, B. Babic, T. Evgeniou, I.G. Cohen, The need for a system view to regulate artificial intelligence/machine learning-based software as medical device, Npj Digit. Med. 3 (1) (2020) 1, http://dx.doi.org/10.1038/s41746-020-0262-2.

[31] J. Vaidya, X. Jiang, A scalable privacy-preserving data generation methodology for exploratory analysis, in: AMIA Annual Symposium Proceedings, 2018, p. 10.

[32] J. Walonoski, et al., Synthea™ novel coronavirus (COVID-19) model and synthetic data set, 2020, p. 8.

[33] B.Z. Harvey, R.T. Sirna, M.B. Houlihan, Learning by design: Hands-on learning, Am. School Board J. 186 (2) (1998) 22–25.

[34] S.K. Helfer, M. Mmel le, F. Bathelt, M. Sedlmayr, Generating enriched synthetic german hospital claims data – a use case driven approach, German Medical Data Sciences: Bringing Data to Life, 2021, pp. 58–65, http://dx.doi.org/10.3233/SHTI210051.

[35] E. Begoli, K. Brown, S. Srinivas, S. Tamang, SynthNotes: A generator framework for high-volume, high-fidelity synthetic mental health notes, in: 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 951–958, http://dx.doi.org/10.1109/BigData.2018.8621981.

[36] Y. Du, S. Lin, Z. Huang, Generation of semantic patient data for depression, in: Health Information Science, 2017, pp. 102–112, http://dx.doi.org/10.1007/978-3-319-69182-4_11.

[37] Y. Jiang, L. Mosquera, B. Jiang, L. Kong, K.E. Emam, Measuring re-identification risk using a synthetic estimator to enable data sharing, PLoS One 17 (6) (2022) e0269097, http://dx.doi.org/10.1371/journal.pone.0269097.

[38] A. Yale, S. Dash, K. Bhanot, I. Guyon, J.S. Erickson, K.P. Bennett, Synthesizing quality open data assets from private health research studies, in: W. Abramowicz, G. Klein (Eds.), Business Information Systems Workshops, 394, Springer International Publishing, Cham, 2020, pp. 324–335, http://dx.doi.org/10.1007/978-3-030-61146-0_26.

[39] L. Wang, W. Zhang, X. He, Continuous patient-centric sequence generation via sequentially coupled adversarial learning, in: Database Systems for Advanced Applications, 2019, pp. 36–52, http://dx.doi.org/10.1007/978-3-030-18579-4_3.

[40] S. Rashidian, et al., SMOOTH-GAN: Towards sharp and smooth synthetic EHR data generation, in: Artificial Intelligence in Medicine, Cham, 2020, pp. 37–48, http://dx.doi.org/10.1007/978-3-030-59137-3_4.

[41] Z. Farou, N. Mouhoub, T. Horváth, Data generation using gene expression generator, in: Intelligent Data Engineering and Automated Learning – IDEAL 2020, Cham, 2020, pp. 54–65, http://dx.doi.org/10.1007/978-3-030-62365-4_6.

[42] Z. Wang, P. Myles, A. Tucker, Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy, Comput. Intell. (2021) coin.12427, http://dx.doi.org/10.1111/coin.12427.

[43] F. Yang, et al., Grouped correlational generative adversarial networks for discrete electronic health records, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, San Diego, CA, USA, 2019, pp. 906–913, http://dx.doi.org/10.1109/BIBM47256.2019.8983215.

[44] A.Y. Ng, M.I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, in: Advances in Neural Information Processing Systems, 2002, pp. 841–848.

[45] G.-F. J, C. E, Synthetic observational health data with GANs: from slow adoption to a boom in medical research and ultimately digital twins? 2020, http://dx.doi.org/10.22541/au.158921777.79483839/v2.

[46] K. El Emam, L. Mosquera, R. Hoptroff, Practical Synthetic Data Generation - Balancing Privacy and the Broad Availability of Data, first ed., O'Reilly, 2020.

[47] K. Malloch, T. Porter-O'Grady, Positive Deviance: Advancing Innovation to Transform Healthcare, Springer Publishing Company, 2020, Accessed: Oct. 25, 2020. [Online]. Available: https://connect.springerpub.com/content/book/978-0-8261-9625-5/part/part02/chapter/ch13.

[48] Donald B. Rubin, Statistical disclosure limitation, J. Off. Stat. 9 (2) (1993) 461–468.

[49] D.S. Dhami, M. Das, S. Natarajan, Knowledge Intensive Learning of Generative Adversarial Networks, San Diego, 2020, p. 6.

[50] National COVID Cohort Collaborative (N3C), National center for advancing translational sciences, 2020, https://ncats.nih.gov/n3c (accessed Jan. 05, 2023).

[51] J.P. Reiter, R. Mitra, Estimating risks of identification disclosure in partially synthetic data, JPC 1 (1) (2009) 1, http://dx.doi.org/10.29012/jpc.v1i1.567.

[52] J.M. Abowd, L. Vilhuber, How protective are synthetic data? in: Privacy in Statistical Databases, Berlin, Heidelberg, 2008, pp. 239–246, http://dx.doi.org/10.1007/978-3-540-87471-3_20.

[53] S. McLachlan, Realism in synthetic data generation, 147.

[54] M. Hittmeir, A. Ekelhart, R. Mayer, On the utility of synthetic data: An empirical evaluation on machine learning tasks, in: Proceedings of the 14th International Conference on Availability, Reliability and Security - ARES '19, Canterbury, CA, United Kingdom, 2019, pp. 1–6, http://dx.doi.org/10.1145/3339252.3339281.

[55] Digital health data: A comprehensive review of privacy and security risks and some recommendations, 2021, https://ibn.idsi.md/vizualizare_articol/46521 (accessed May 20, 2021).

[56] 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 (11) (2020) e23139, http://dx.doi.org/10.2196/23139.
[57] 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.
[58] S.I. Nikolenko, Privacy guarantees in synthetic data, in: S.I. Nikolenko (Ed.), Synthetic Data for Deep Learning, Springer International Publishing, Cham, 2021, pp. 269–283, http://dx.doi.org/10.1007/978-3-030-75178-4_11.
[59] J. Jordon, A. Wilson, M. van der Schaar, Synthetic data: Opening the data floodgates to enable faster, more directed development of machine learning methods, 2020, arXiv. Accessed: Aug. 10, 2022. [Online]. Available: http://arxiv.org/abs/2012.04580.
[60] J. Jordon, et al., Synthetic data – what, why and how? 2022, http://dx.doi.org/10.48550/arXiv.2205.03257, arXiv.
[61] J. Coutinho-Almeida, P.P. Rodrigues, R.J. Cruz-Correia, GANs for tabular healthcare data generation: A review on utility and privacy, in: Discovery Science, Cham, 2021, pp. 282–291, http://dx.doi.org/10.1007/978-3-030-88942-5_22.
[62] G. Ghosheh, J. Li, T. Zhu, A review of generative adversarial networks for electronic health records: applications, evaluation measures and data sources, 2022, arXiv. Accessed: Jan. 02, 2023. [Online]. Available: http://arxiv.org/abs/2203.07018.
[63] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy, SP, 2017, pp. 3–18, http://dx.doi.org/10.1109/SP.2017.41.
[64] B. Kitchenham, P. Brereton, A systematic review of systematic review process research in software engineering, Inf. Softw. Technol. 55 (12) (2013) 2049–2075, http://dx.doi.org/10.1016/j.infsof.2013.07.010.
[65] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering – A systematic literature review, Inf. Softw. Technol. 51 (1) (2009) 7–15, http://dx.doi.org/10.1016/j.infsof.2008.09.009.
[66] T. Wang, et al., A review on medical imaging synthesis using deep learning and its clinical applications, J. Appl. Clin. Med. Phys. 22 (1) (2021) 11–36, http://dx.doi.org/10.1002/acm2.13121.
[67] N. Ruiz, K. Muralidhar, J. Domingo-Ferrer, On the privacy guarantees of synthetic data: A reassessment from the maximum-knowledge attacker perspective, in: Privacy in Statistical Databases, Cham, 2018, pp. 59–74, http://dx.doi.org/10.1007/978-3-319-99771-1_5.
[68] Z. Huang, F. van Harmelen, A. ten Teije, K. Dentler, Knowledge-based patient data generation, in: D. Riaño, R. Lenz, S. Miksch, M. Peleg, M. Reichert, A. ten Teije (Eds.), Process Support and Knowledge Representation in Health Care, Vol. 8268, Springer International Publishing, Cham, 2013, pp. 83–96, http://dx.doi.org/10.1007/978-3-319-03916-9_7.
[69] S. McLachlan, K. Dube, T. Gallagher, J.A. Simmonds, N. Fenton, Realistic synthetic data generation: The ATEN framework, in: A. Cliquet, S. Wiebe, P. Anderson, G. Saggio, R. Zwiggelaar, H. Gamboa, A. Fred, S. Bermúdez i Badia (Eds.), Biomedical Engineering Systems and Technologies, Vol. 1024, Springer International Publishing, Cham, 2019, pp. 497–523, http://dx.doi.org/10.1007/978-3-030-29196-9_25.
[70] S. McLachlan, K. Dube, T. Gallagher, Using the CareMap with health incidents statistics for generating the realistic synthetic electronic healthcare record, in: 2016 IEEE International Conference on Healthcare Informatics, ICHI, Chicago, IL, USA, 2016, pp. 439–448, http://dx.doi.org/10.1109/ICHI.2016.83.
[71] J. Walonoski, et al., Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, J. Am. Med. Inform. Assoc. 25 (3) (2018) 230–238, http://dx.doi.org/10.1093/jamia/ocx079.
[72] J. Chen, D. Chun, M. Patel, E. Chiang, J. James, The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures, BMC Med. Inform. Decis. Mak. 19 (1) (2019) 44, http://dx.doi.org/10.1186/s12911-019-0793-0.
[73] A. Oganian, V-dispersed synthetic data based on a mixture model with constraints, in: J. Domingo-Ferrer (Ed.), Privacy in Statistical Databases, Vol. 8744, Springer International Publishing, Cham, 2014, pp. 200–212, http://dx.doi.org/10.1007/978-3-319-11257-2_16.
[74] A. Oganian, J. Domingo-Ferrer, Local synthesis for disclosure limitation that satisfies probabilistic k-anonymity criterion, 2019, p. 28.
[75] J. Zhang, G. Cormode, C.M. Procopiuc, D. Srivastava, X. Xiao, PrivBayes: Private data release via Bayesian networks, ACM Trans. Database Syst. 42 (4) (2017) 1–41, http://dx.doi.org/10.1145/3134428.
[76] Y. Park, J. Ghosh, M. Shankar, Perturbed gibbs samplers for generating large-scale privacy-safe synthetic health data, in: 2013 IEEE International Conference on Healthcare Informatics, Philadelphia, PA, USA, 2013, pp. 493–498, http://dx.doi.org/10.1109/ICHI.2013.76.
[77] Y. Park, J. Ghosh, Pegs: Perturbed gibbs samplers that gen- erate privacy-compliant synthetic data, 2014, p. 30.
[78] R. Foraker, A. Guo, J. Thomas, N. Zamstein, P.R. Payne, A. Wilcox, The national COVID cohort collaborative: Analyses of original and computationally derived electronic health record data, J. Med. Internet Res. 23 (10) (2021) e30697, http://dx.doi.org/10.2196/30697.
[79] J.A. Thomas, R.E. Foraker, N. Zamstein, P.R.O. 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. Assoc. 29 (8) (2022) http://dx.doi.org/10.1101/2021.07.06.21259051.
[80] D. Kaur, et al., Application of Bayesian networks to generate synthetic health data, J. Am. Med. Inform. Assoc. 28 (4) (2021) 801–811, http://dx.doi.org/10.1093/jamia/ocaa303.
[81] E. Choi, S. Biswal, B. Malin, J. Duke, W.F. Stewart, J. Sun, Generating multi-label discrete patient records using generative adversarial networks, in: Machine Learning for Healthcare Conference, 2017, pp. 286–305, Accessed: May 10, 2021. [Online]. Available: http://proceedings.mlr.press/v68/choi17a.html.
[82] P. Jackson, M. Lussetti, Extending a generative adversarial network to produce medical records with demographic characteristics and health system use, in: 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON, Vancouver, BC, Canada, 2019, pp. 0515–0518, http://dx.doi.org/10.1109/IEMCON.2019.8936168.
[83] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, Y. Kim, Data synthesis based on generative adversarial networks, Proc. VLDB Endow. 11 (10) (2018) 1071–1083, http://dx.doi.org/10.14778/3231751.3231757.
[84] M.L. Fang, D.S. Dhami, K. Kersting, DP-CTGAN: Differentially private medical data generation using CTGANs, in: M. Michalowski, S.S.R. Abidi, S. Abidi (Eds.), Artificial Intelligence in Medicine, 13263, Springer International Publishing, Cham, 2022, pp. 178–188, http://dx.doi.org/10.1007/978-3-031-09342-5_17.
[85] E.B. Ozyigit, T.N. Arvanitis, G. Despotou, Generation of realistic synthetic validation healthcare datasets using generative adversarial networks, 2020, p. 4.
[86] M.K. Baowaly, C.-L. Liu, K.-T. Chen, Realistic data synthesis using enhanced generative adversarial networks, in: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering, AIKE, Sardinia, Italy, 2019, pp. 289–292, http://dx.doi.org/10.1109/AIKE.2019.00057.
[87] M.K. Baowaly, C.-C. Lin, C.-L. Liu, K.-T. Chen, Synthesizing electronic health records using improved generative adversarial networks, J. Am. Med. Inform. Assoc. 26 (3) (2019) 228–241, http://dx.doi.org/10.1093/jamia/ocy142.
[88] A. Torfi, E.A. Fox, Corgan: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records, 2020, arXiv:2001.09346 [cs, stat]. Accessed: Sep. 14, 2020. [Online]. Available: http://arxiv.org/abs/2001.09346.
[89] D. Hazra, Y.-C. Byun, SynSigGAN: Generative adversarial networks for synthetic biomedical signal generation, Biology 9 (12) (2020) 441, http://dx.doi.org/10.3390/biology9120441.
[90] Z. Zhang, C. Yan, D.A. Mesa, J. Sun, B.A. Malin, Ensuring electronic medical record simulation through better training, modeling, and evaluation, J. Am. Med. Inform. Assoc. 27 (1) (2020) 99–108, http://dx.doi.org/10.1093/jamia/ocz161.
[91] C. Yan, Z. Zhang, S. Nyemba, B.A. Malin, Generating electronic health records with multiple data types and constraints, in: AMIA Annu Symp Proc, Vol. 2020, 2021, pp. 1335–1344.
[92] J. Guan, R. Li, S. Yu, X. Zhang, A method for generating synthetic electronic medical record text, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (1) (2021) 173–182, http://dx.doi.org/10.1109/TCBB.2019.2948985.
[93] R. Venugopal, N. Shafqat, I. Venugopal, B.M.J. Tillbury, H.D. Stafford, A. Bourazeri, Privacy preserving generative adversarial networks to model electronic health records, Neural Netw. 153 (2022) 339–348, http://dx.doi.org/10.1016/j.neunet.2022.06.022.
[94] S. Dash, A. Yale, I. Guyon, K.P. Bennett, Medical time-series data generation using generative adversarial networks, in: Artificial Intelligence in Medicine, Cham, 2020, pp. 382–391.
[95] K. Bhanot, J. Pedersen, I. Guyon, K.P. Bennett, Investigating synthetic medical time-series resemblance, Neurocomputing 494 (2022) 368–378, http://dx.doi.org/10.1016/j.neucom.2022.04.097.
[96] K. Bhanot, S. Dash, J. Pedersen, I. Guyon, K. Bennett, Quantifying resemblance of synthetic medical time-series, in: ESANN 2021 Proceedings, Online event (Bruges, Belgium), 2021, pp. 611–616, http://dx.doi.org/10.14428/esann/2021.ES2021-108.
[97] J. Jordon, J. Yoon, PATE-GAN: Generating synthetic data with differential private guarantees, in: ICLR 2019, 2019, p. 21.

[98] Y. Liu, J. Peng, J.J.Q. Yu, Y. Wu, PPGAN: Privacy-preserving generative adversarial network, in: 2019 IEEE 25th International Conference on Parallel and Distributed Systems, ICPADS, 2019, pp. 985–989, http://dx.doi.org/10.1109/ICPADS47876.2019.00150.

[99] B.K. Beaulieu-Jones, et al., Privacy-preserving generative deep neural networks support clinical data sharing, Circ: Cardiovasc. Qual. Outcomes 12 (7) (2019) http://dx.doi.org/10.1161/CIRCOUTCOMES.118.005122.

[100] S. Wang, C. Rudolph, S. Nepal, M. Grobler, S. Chen, PART-GAN: Privacy-preserving time-series sharing, in: Artificial Neural Networks and Machine Learning – ICANN 2020, Cham, 2020, pp. 578–593, http://dx.doi.org/10.1007/978-3-030-61609-0_46.

[101] J. Yoon, L.N. Drumright, M. van der Schaar, Anonymization through data synthesis using generative adversarial networks (ADS-GAN), IEEE J. Biomed. Health Inform. 24 (8) (2020) 2378–2388, http://dx.doi.org/10.1109/JBHI.2020.2980262.

[102] S. Imtiaz, M. Arsalan, V. Vlassov, R. Sadre, Synthetic and private smart health care data generation using GANs, in: 2021 International Conference on Computer Communications and Networks, ICCCN, Athens, Greece, 2021, pp. 1–7, http://dx.doi.org/10.1109/ICCCN52240.2021.9522203.

[103] D. Lee, et al., Generating sequential electronic health records using dual adversarial autoencoder, J. Am. Med. Inform. Assoc. 27 (9) (2020) 1411–1419, http://dx.doi.org/10.1093/jamia/ocaa119.

[104] A. Torfi, E.A. Fox, C.K. Reddy, Differentially private synthetic medical data generation using convolutional GANs, Inform. Sci. 586 (2022) 485–500, http://dx.doi.org/10.1016/j.ins.2021.12.018.

[105] Z. Zhang, C. Yan, T.A. Lasko, J. Sun, B.A. Malin, Synteg: a framework for temporal structured electronic health data simulation, J. Am. Med. Inform. Assoc. 28 (3) (2021) 596–604, http://dx.doi.org/10.1093/jamia/ocaa262.

[106] Z. Zhang, C. Yan, B.A. Malin, Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation, J. Am. Med. Inform. Assoc. 29 (11) (2022) 1890–1898, http://dx.doi.org/10.1093/jamia/ocac131.

[107] A. Hernandez-Matamoros, H. Fujita, H. Perez-Meana, A novel approach to create synthetic biomedical signals using BiRNN, Inform. Sci. 541 (2020) 218–241, http://dx.doi.org/10.1016/j.ins.2020.06.019.

[108] N.C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, L. Sweeney, Privacy preserving synthetic data release using deep learning, in: M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, G. Ifrim (Eds.), Machine Learning and Knowledge Discovery in Databases, 11051, Springer International Publishing, Cham, 2019, pp. 510–526, http://dx.doi.org/10.1007/978-3-030-10925-7_31.

[109] S. Biswal, et al., EVA: Generating longitudinal electronic health records using conditional variational autoencoders, in: Proceedings of the 6th Machine Learning for Healthcare Conference, 2021, pp. 260–282, Accessed: Aug. 23, 2022. [Online]. Available: https://proceedings.mlr.press/v149/biswal21a.html.

[110] K.E. Emam, L. Mosquera, C. Zheng, Optimizing the synthesis of clinical trial data using sequential trees, J. Am. Med. Inform. Assoc. 28 (1) (2021) 3–13, http://dx.doi.org/10.1093/jamia/ocaa249.

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

[112] D. Dua, C. Graff, UCI machine learning repository: Data sets, 2021, http://archive.ics.uci.edu/ml/datasets.php (accessed May 20, 2021).

[113] SEER data & software, SEER, 2021, https://seer.cancer.gov/data-software/index.html (accessed Jun. 20, 2021).

[114] Cervical cancer dataset, 2022, https://www.kaggle.com/dataset/7ce132ff89fb98573a1501864de17d1f8f928dee7034296be3fd13d7a59bd710 (accessed May 25, 2022).

[115] E. Brophy, Z. Wang, Q. She, T. Ward, Generative adversarial networks in time series: A survey and taxonomy, 2021, arXiv:2107.11098 [cs]. Accessed: Aug. 23, 2021. [Online]. Available: http://arxiv.org/abs/2107.11098.

[116] A.E.W. Johnson, et al., MIMIC-III, a freely accessible critical care database, Sci Data 3 (1) (2016) 160035, http://dx.doi.org/10.1038/sdata.2016.35.

[117] D. Saxena, J. Cao, Generative adversarial networks (GANs): Challenges, solutions, and future directions, 42.

[118] H. Quick, L.A. Waller, Using spatiotemporal models to generate synthetic data for public use, Spat. Spat.-Temp. Epidemiol. 27 (2018) 37–45, http://dx.doi.org/10.1016/j.sste.2018.08.004.

[119] K. Dube, T. Gallagher, Approach and method for generating realistic synthetic electronic healthcare records for secondary use, in: J. Gibbons, W. MacCaull (Eds.), Foundations of Health Information Engineering and Systems, Vol. 8315, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, pp. 69–86, http://dx.doi.org/10.1007/978-3-642-53956-5_6.

[120] Khaled El Emam, Richard Hoptroff, The synthetic data paradigm for using and sharing data, Cutter Executive Update 19 (6) (2019).

[121] P. Kumar, M. Shoukri, Copula functions for modelling dependence structure with applications in the analysis of clinical data, J. Indian Soc. Agric. Statist. 61 (2007).

[122] MDClone - the world's most powerful healthcare data platform, MDClone, 2023, https://www.mdclone.com (accessed Jan. 03, 2023).

[123] Generative adversarial nets | proceedings of the 27th international conference on neural information processing systems - volume 2, 2021, https://dl.acm.org/doi/abs/10.5555/2969033.2969125 (accessed Feb. 10, 2021).

[124] L. Lan, et al., Generative adversarial networks and its applications in biomedical informatics, Front. Public Health 8 (2020) http://dx.doi.org/10.3389/fpubh.2020.00164.

[125] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: A review, Med. Image Anal. 58 (2019) 101552, http://dx.doi.org/10.1016/j.media.2019.101552.

[126] L. Yu, W. Zhang, J. Wang, Y. Yu, SeqGAN: sequence generative adversarial nets with policy gradient, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 2017, pp. 2852–2858.

[127] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, 2014, arXiv:1312.6114 [cs, stat]. Accessed: Mar. 17, 2021. [Online]. Available: http://arxiv.org/abs/1312.6114.

[128] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507, http://dx.doi.org/10.1126/science.1127647.

[129] P. Jackson, M. Lussetti, Extending a generative adversarial network to produce medical records with demographic characteristics and health system use, in: 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON, Vancouver, BC, Canada, 2019, pp. 0515–0518, http://dx.doi.org/10.1109/IEMCON.2019.8936168.

[130] R.D. Hjelm, A.P. Jacob, T. Che, A. Trischler, K. Cho, Y. Bengio, Boundary-seeking generative adversarial networks, in: Presented at the 6th International Conference on Learning Representations, ICLR 2018, 2018, Accessed: Mar. 18, 2021. [Online]. Available: https://nyuscholars.nyu.edu/en/publications/boundary-seeking-generative-adversarial-networks.

[131] Improved training of wasserstein GANs | proceedings of the 31st international conference on neural information processing systems, 2021, https://dl.acm.org/doi/abs/10.5555/3295222.3295327 (accessed Mar. 18, 2021).

[132] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, arXiv:1411.1784 [cs, stat]. Accessed: Sep. 11, 2020. [Online]. Available: http://arxiv.org/abs/1411.1784.

[133] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, K.P. Bennett, Assessing privacy and quality of synthetic health data, in: Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse, Pittsburgh Pennsylvania, 2019, pp. 1–4, http://dx.doi.org/10.1145/3359115.3359124.

[134] M. Gong, Y. Xie, K. Pan, K. Feng, A.K. Qin, A survey on differentially private machine learning [review article], IEEE Comput. Intell. Mag. 15 (2) (2020) 49–64, http://dx.doi.org/10.1109/MCI.2020.2976185.

[135] J. Yoon, End-to-end machine learning frameworks for medicine: Data imputation, model interpretation and synthetic data generation, 2020, p. 168.

[136] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, K. Talwar, Semi-supervised knowledge transfer for deep learning from private training data, 2017, arXiv:1610.05755 [cs, stat]. Accessed: Sep. 02, 2020. [Online]. Available: http://arxiv.org/abs/1610.05755.

[137] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, Ú. Erlingsson, Scalable private learning with pate, 2018, p. 34.

[138] Y. Long, S. Lin, Z. Yang, C.A. Gunter, B. Li, Scalable differentially private generative student model via PATE, 2019, arXiv:1906.09338 [cs, stat]. Accessed: Sep. 03, 2020. [Online]. Available: http://arxiv.org/abs/1906.09338.

[139] L. Xie, K. Lin, S. Wang, F. Wang, J. Zhou, Differentially private generative adversarial network, 2018, arXiv:1802.06739 [cs, stat]. Accessed: Apr. 01, 2021. [Online]. Available: http://arxiv.org/abs/1802.06739.

[140] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional GAN, 2019, arXiv. Accessed: Aug. 23, 2022. [Online]. Available: http://arxiv.org/abs/1907.00503.

[141] M. Ranzato, S. Chopra, M. Auli, W. Zaremba, Sequence level training with recurrent neural networks: 4th international conference on learning representations, in: ICLR 2016, 2016, Accessed: Jan. 08, 2023. [Online]. Available: http://www.scopus.com/inward/record.url?scp=85083951479&partnerID=8YFLogxK.

[142] Z. Shi, L. He, Application of neural networks in medical image processing, 4.

[143] Z. Wang, P. Myles, A. Tucker, Generating and evaluating synthetic UK primary care data: Preserving data utility & patient privacy, in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems, CBMS, Cordoba, Spain, 2019, pp. 126–131, http://dx.doi.org/10.1109/CBMS.2019.00036.

[144] D. Riaño, A. Fernández-Pérez, Simulation-based episodes of care data synthetization for chronic disease patients, Knowl. Represent. Health Care (2016) 36–50, http://dx.doi.org/10.1007/978-3-319-55014-5_3.

[145] S. Schiff, M. Gehrke, R. Möller, Efficient enriching of synthesized relational patient data with time series data, Procedia Comput. Sci. 141 (2018) 531–538, http://dx.doi.org/10.1016/j.procs.2018.10.130.
[146] X. Larrea, et al., Synthetic subject generation with coupled coherent time series data, Eng. Proc. 18 (1) (2022) 1, http://dx.doi.org/10.3390/engproc2022018007.
[147] 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, 2020, pp. 38–47.
[148] C.A. Libbi, J. Trienes, D. Trieschnigg, C. Seifert, Generating synthetic training data for supervised de-identification of electronic health records, Future Internet 13 (5) (2021) 5, http://dx.doi.org/10.3390/fi13050136.
[149] K. Bhanot, S. Dash, J. Pedersen, I. Guyon, K. Bennett, Quantifying resemblance of synthetic medical time-series, in: ESANN 2021 Proceedings, Online event (Bruges, Belgium), 2021, pp. 611–616, http://dx.doi.org/10.14428/esann/2021.ES2021-108.
[150] S. Biswal, S. Ghosh, EVA: Generating longitudinal electronic health records using conditional variational autoencoders, 22.
[151] The synthetic data vault. Put synthetic data to work! 2022, https://sdv.dev/ (accessed Aug. 29, 2022).
[152] J. Jordon, et al., Synthetic data – what, why and how? 2022, arXiv. Accessed: Aug. 09, 2022. [Online]. Available: http://arxiv.org/abs/2205.03257.
[153] S.L. Hyland, C. Esteban, G. Rätsch, Real-valued (medical) time series generation with recurrent conditional GANs, 12.
[154] J. Jordon, J. Yoon, M. van der Schaar, Measuring the quality of synthetic data for use in competitions, 2018, arXiv:1806.11345 [cs, stat]. Accessed: Sep. 03, 2020. [Online]. Available: http://arxiv.org/abs/1806.11345.
[155] O. Mendelevitch, M.D. Lesh, Fidelity and privacy of synthetic medical data, 2021, arXiv:2101.08658 [cs]. Accessed: Jul. 28, 2021. [Online]. Available: http://arxiv.org/abs/2101.08658.
[156] M. Hittmeir, R. Mayer, A. Ekelhart, A baseline for attribute disclosure risk in synthetic data, in: Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy, New Orleans LA USA, 2020, pp. 133–143, http://dx.doi.org/10.1145/3374664.3375722.
[157] K. El Emam, L. Mosquera, X. Fang, Validating a membership disclosure metric for synthetic health data, JAMIA Open 5 (4) (2022) ooac083, http://dx.doi.org/10.1093/jamiaopen/ooac083.
[158] D. Chen, N. Yu, Y. Zhang, M. Fritz, GAN-leaks: A taxonomy of membership inference attacks against generative models, in: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event USA, 2020, pp. 343–362, http://dx.doi.org/10.1145/3372297.3417238.
[159] Z. Zhang, C. Yan, B.A. Malin, Membership inference attacks against synthetic health data, J. Biomed. Inform. 125 (2022) 103977, http://dx.doi.org/10.1016/j.jbi.2021.103977.
[160] K.V. Saboo, A. Choudhary, Y. Cao, G.A. Worrell, D.T. Jones, R.K. Iyer, Reinforcement learning based disease progression model for Alzheimer's disease, 13.
[161] D. Monterde, E. Vela, M. Clèries, L. Garcia-Eroles, J. Roca, P. Pérez-Sust, Multimorbidity as a predictor of health service utilization in primary care: a registry-based study of the Catalan population, BMC Fam. Pract. 21 (1) (2020) 39, http://dx.doi.org/10.1186/s12875-020-01104-1.
[162] K. El Emam, L. Mosquera, X. Fang, A. El-Hussuna, Utility metrics for evaluating synthetic health data generation methods: Validation study, JMIR Med. Inform. 10 (4) (2022) e35734, http://dx.doi.org/10.2196/35734.
[163] C. Yan, et al., A multifaceted benchmarking of synthetic electronic health record generation models, Nature Commun. 13 (1) (2022) 7609, http://dx.doi.org/10.1038/s41467-022-35295-1.
[164] UNOS dataset, 2021, https://unos.org/data/ (accessed Jun. 23, 2021).
[165] BSA inpatient claims PUF | CMS, 2021, https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/BSAPUFS/Inpatient_Claims (accessed Jun. 20, 2021).
[166] National syndromic surveillance program (NSSP) | CDC, 2022, https://www.cdc.gov/nssp/index.html (accessed Aug. 11, 2022).
[167] Project data sphere® data platform | project data sphere, 2021, https://www.projectdatasphere.org/data-platform (accessed Jun. 20, 2021).
[168] .
[169] S.J. Pocock, et al., Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies, Eur. Heart J. 34 (19) (2013) 1404–1413, http://dx.doi.org/10.1093/eurheartj/ehs337.
[170] K. Fernandes, J.S. Cardoso, J. Fernandes, Transfer learning with partial observability applied to cervical cancer screening, in: Pattern Recognition and Image Analysis, Cham, 2017, pp. 243–250, http://dx.doi.org/10.1007/978-3-319-58838-4_27.
[171] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, R.S. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in: Proc Annu Symp Comput Appl Med Care, 1988, pp. 261–265.
[172] A. Tsanas, M. Little, P. McSharry, L. Ramig, Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests, Nat. Prec. (2009) 1, http://dx.doi.org/10.1038/npre.2009.3920.1.
[173] R.G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, C.E. Elger, Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state, Phys. Rev. E 64 (6 Pt 1) (2001) 061907, http://dx.doi.org/10.1103/PhysRevE.64.061907.
[174] L. Lin, C. Warren-Gash, L. Smeeth, P.-C. Chen, Data resource profile: the national health insurance research database (NHIRD), Epidemiol. Health 40 (2018) e2018062, http://dx.doi.org/10.4178/epih.e2018062.
[175] G.B. Moody, R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. Mag. 20 (3) (2001) 45–50, http://dx.doi.org/10.1109/51.932724.
[176] G.B. Goldberger, A.L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, R. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation (2000) http://dx.doi.org/10.13026/C2F305, [Online]. physionet.org.
[177] M.A.F. Pimentel, et al., Toward a robust estimation of respiratory rate from pulse oximeters, IEEE Trans. Biomed. Eng. 64 (8) (2017) 1914–1923, http://dx.doi.org/10.1109/TBME.2016.2613124.
[178] Detti, Paolo, Siena scalp EEG database. PhysioNet. http://dx.doi.org/10.13026/5D4A-J060.
[179] P. Detti, G. Vatti, G. Zabalo Manrique de Lara, EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings, Processes 8 (7) (2020) 7, http://dx.doi.org/10.3390/pr8070846.
[180] MIMIC Database, MIMIC, 2021, https://mimic.mit.edu/ (accessed May 20, 2021).
[181] Synthetic derivative | department of biomedical informatics, 2021, https://www.vumc.org/dbmi/synthetic-derivative (accessed Jun. 20, 2021).
[182] Patient Discharge Dataset, Office of Statewide Health Planning and Development (OSHPD), 2014, 2009. Accessed: Jun. 20, 2021. [Online]. Available: https://oshpd.ca.gov/visualizations/patient-discharge-data-by-principal-diagnosis-group-2009-2014/.
[183] Clinical practice research datalink | CPRD, 2021, https://www.cprd.com/ (accessed May 20, 2021).
[184] S. CTSI, Cerner Health Facts, SC CTSI, 2021, https://sc-ctsi.org/resources/cerner-health-facts (accessed May 20, 2021).
[185] D. Mongin, J. García Romero, J.R. Alvero Cruz, Treadmill Maximal Exercise Tests from the Exercise Physiology and Human Performance Lab of the University of Malaga (Version 1.0.1), PhysioNet, 2021, http://dx.doi.org/10.13026/7ezk-j442.
[186] D. Mongin, C. Chabert, D.S. Courvoisier, J. García-Romero, J.R. Alvero-Cruz, Heart rate recovery to assess fitness: comparison of different calculation methods in a large cross-sectional study, Res. Sports Med. (2021) 1–14, http://dx.doi.org/10.1080/15438627.2021.1954513.
[187] VPS PICU | virtual pediatric systems, 2023, https://myvps.org/vps-picu/ (accessed Jan. 04, 2023).
[188] D. Kaur, et al., Application of Bayesian networks to generate synthetic health data, J. Am. Med. Inform. Assoc. (ocaa303) (2020) http://dx.doi.org/10.1093/jamia/ocaa303.
[189] COVID-19 Canada, 2022, https://resources-covid19canada.hub.arcgis.com/ (accessed Aug. 24, 2022).