Annual Review of Biomedical Data Science

Conditional Generative Models for Synthetic Tabular Data: Applications for Precision Medicine and Diverse Representations

Kara Liu[1] and Russ B. Altman[2]

[1]Department of Computer Science, Stanford University, Stanford, California, USA; email: karaliu@stanford.edu

[2]Departments of Bioengineering and Genetics, Stanford University, Stanford, California, USA

Annu. Rev. Biomed. Data Sci. 2025. 8:21–49

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-103123-094844

Keywords

generative models, synthetic data, tabular data, precision medicine, fairness, diversity, healthcare, EHRs

Abstract

Tabular medical datasets, like electronic health records (EHRs), biobanks, and structured clinical trial data, are rich sources of information with the potential to advance precision medicine and optimize patient care. However, real-world medical datasets have limited patient diversity and cannot simulate hypothetical outcomes, both of which are necessary for equitable and effective medical research. Fueled by recent advancements in machine learning, generative models offer a promising solution to these data limitations by generating enhanced synthetic data. This review highlights the potential of conditional generative models (CGMs) to create patient-specific synthetic data for a variety of precision medicine applications. We survey CGM approaches that tackle two medical applications: correcting for data representation biases and simulating digital health twins. We additionally explore how the surveyed methods handle modeling tabular medical data and briefly discuss evaluation criteria. Finally, we summarize the technical, medical, and ethical challenges that must be addressed before CGMs can be effectively and safely deployed in the medical field.

1. INTRODUCTION

The field of precision medicine, defined as the personalization of medical treatment based on each patient's unique phenotype, has recently been fueled by advancements in machine learning (ML) and genomic sequencing, as well as the proliferation of tabular medical datasets (1). In particular, tabular datasets, like electronic health records (EHRs) and structured clinical trial data, provide a wealth of information and enable the study of how medical decisions dynamically impact a patient or cohort. However, precision medicine requires diverse and simulatable data, qualities currently lacking in real-world datasets. With respect to diversity, medical datasets have historically under- or misrepresented certain health conditions and demographic groups (2, 3), creating blind spots that clinicians are ill-equipped to treat. In addition, existing datasets only capture observed patient outcomes, whereas the ability to simulate unobserved trajectories could enhance treatment optimization and prevent adverse outcomes by forecasting future states.

The ideal solution to overcome existing limitations in medical datasets is to gather data covering every treatment, phenotype, and patient cohort necessary for creating personalized treatment plans. However, such data collection is currently infeasible, requiring a level of randomization, organization, financial cost, and social harmony [e.g., facilitating trust with marginalized groups (4)] that the healthcare sector is currently incapable of. Until all-encompassing data collection becomes a reality, it would benefit medical research to augment real-world tabular datasets with equitable patient representation and the capacity to simulate patient trajectories.

One promising solution to flawed data is the application of generative models for synthetic tabular data. Broadly, generative models use statistical methods like ML to learn the data distribution and sample new data points. Recently, generative ML methods like generative adversarial networks (GANs) (5) and denoising diffusion probabilistic models (DDPMs) (6) have proven effective in generating high-quality synthetic medical image (7, 8) and EHR data (9).[1]

However, standard generative models are limited to learning, and thus sampling from, the full data distribution. In this review, we highlight the potential of conditional generative models (CGMs), whereby an ML model learns the conditional probability distribution of the data conditioned on user-specified patient characteristics. Because they enable the sampling of patient-specific data, CGMs are a promising technology for precision medicine that can generate patient trajectories, optimize treatment by simulating intervention effect, and ensure fair representation of marginalized cohorts.

This review surveys the recent development of CGMs for tabular data from the lens of medical applications. We structure the review as follows:

- In Section 2, we cover important definitions, discuss the state of the field of generative models, and outline existing applications to the medical domain.
- In Section 3, we define the review's scope and inclusion/exclusion criteria, which identified 43 relevant works.
- In Section 4, we highlight and discuss two promising applications of CGMs to medicine: (*a*) correcting for biases in real datasets and (*b*) simulating hypothetical patient-specific outcomes. We also review how these approaches tackle the technical challenges in modeling tabular medical data and conclude by briefly discussing CGM evaluation.
- In Section 5, we summarize the technical, medical, and ethical challenges that necessitate further exploration.

[1]Henceforth, we use the term EHR to refer to the tabular portion of health record data.

2. BACKGROUND

2.1. Key Definitions

We next outline the problem formulation, provide two key definitions of CGMs, and define data bias.

2.1.1. Problem formulation. Tabular data are defined as mixed-type data containing binary, categorical, integer-valued, and continuous variables. Tabular medical data can be either longitudinal[2] (where every timestep is a single interaction the patient has with the healthcare system, henceforth referred to as a visit) or static (representing a patient at a fixed point in time).

For a patient i and timestamp t, we represent their visit information $v_t^{(i)}$ as an observed collection of K continuous and categorical random variables $v_t^{(i)} = \{x_{t,1}^{(i)}, \dots x_{t,K}^{(i)}\}$. We allow for a continuous variable to be either observed $[x_{t,j}^{(i)} \in \mathbb{R}]$ or unobserved $[x_{t,j}^{(i)} = \emptyset]$. Similarly, any categorical variable with cardinality M can be either observed $[x_{t,j}^{(i)} \in \{c_1, c_2, \dots c_M\}]$ or unobserved $[x_{t,j}^{(i)} = \emptyset]$. We define a data point $s^{(i)}$ as a patient's full health history, represented as a collection of T visits $s^{(i)} = \{v_{t_1}^{(i)}, \dots v_{t_T}^{(i)}\}$ over timestamps $t_1, \dots t_T$. This representation reduces to static data if there is only one timestamp ($T = 1$). We define a regularly sampled time series as one where timestamps are discrete and evenly spaced (e.g., $t_i = i$) and an irregularly sampled time series as one where the timestamps are continuous ($t_i \in \mathbb{R}, t_i < t_{i+1}$).

Given a dataset of N patient trajectories $D = \{s^{(1)}, s^{(2)}, \dots s^{(N)}\}$, a generative model q_ϕ learns the joint distribution over all K variables and T timestamps: $q_\phi(\cup_{j=1}^{T} \cup_{k=1}^{K} x_{t_j,k}) = q_\phi(\cup_{j=1}^{T} v_{t_j})$.

Furthermore, we can define c as the features to condition that are a subset of the $K \cdot T$ patient variables: $c \subset (\cup_{j=1}^{T} \cup_{k=1}^{K} x_{t_j,k})$. Let the function f be a simple postprocessing equation designed to create the final patient trajectory after sampling from the conditional model. We then provide the following definition.

Definition 1. We define a conditional generative model (CGM) q_θ as a model that, given features c, learns the conditional data distribution for the remaining variables $\cup_{j=1}^{T} v_{t_j} - c$ (where $-$ denotes the set difference to exclude regenerating c). A synthetic patient s can then be sampled from $s \sim f(q_\theta(\cup_{j=1}^{T} v_{t_j} - c|c))$.

Explicitly,

1. For static data, c is a subset of $k \ll K$ variables and a CGM will generate the remaining $K - k$ variables $v \sim q_\theta(\cup_{k=1}^{K} x_k - c|c)$. Then the postprocessing function f will concatenate the sample v with c to get the synthetic patient data point $s = (v, c)$.
2. For regular time series, define optional fixed patient attributes as the set of variables u. To generate the next visit v_t at timestep t, the CGM conditions on both fixed patient attributes u and all K variables over the $t - 1$ prior visits: $c = ((\cup_{j=1}^{t-1} v_j), u)$. The next visit can be sampled from $v_t \sim q_\theta(v_t|c) = q_\theta(v_t|(\cup_{j=1}^{t-1} v_j), u)$. If this generation is autoregressively repeated for some n more timesteps, then f can concatenate these n visit samples either with all of the patient's real history to get the semisynthetic patient trajectory $s = (c, v_t, \dots v_{t+n}) = (u, v_1, \dots v_t, \dots v_{t+n})$ or with only u to get the fully synthetic trajectory $s = (u, v_t, \dots v_{t+n})$.
3. For irregular time series, timesteps are no longer discrete and thus must be explicitly represented as a continuous variable. To generate the ith visit, the CGM is conditioned on fixed patient attributes u, all variables from prior visits, and all past timestamp variables:

[2] We use the terms longitudinal and time series interchangeably.

$c = ((\cup_{j=1}^{i-1} v_{t_j}, t_j), u)$. The next visit and its timestamp can be sampled from $v_{t_i} \sim q_\theta(v_{t_i}, t_i|c) = q_\theta(v_{t_i}, t_i|(\cup_{j=1}^{i-1} v_{t_j}, t_j), u)$. A synthetic patient trajectory s can be sampled similarly to the case of regular time series.

Finally, to avoid the reduction of a CGM to a discriminative model like a classifier, we reiterate the focus on data generation and add the additional, albeit subjective, definition.

Definition 2. A CGM should be capable of sampling a fully synthetic dataset of similar dimension (e.g., roughly K variables per timestep) and utility as the real data.[3]

For example, a model limited to next-step time-series forecasting for a single variable is a discriminative predictor and not a CGM. However, an autoregressive model validated on multistep forecasting for all K variables would be considered a CGM. We note that this distinction between discriminative methods and CGMs is imprecise and could be open to interpretation.

2.1.2. Data bias. We represent the real (potentially flawed) data distribution as p, an external ideal distribution as p^*, the learned CGM as q_θ, marginalized group membership as the variable $A = a$, and the rest of the data (all variables excluding A) as X.

Underrepresentation bias occurs when the real data's marginal distribution $p(a)$ does not match the ideal $p^*(a)$. We say q_θ corrects for underrepresentation bias if it matches the ideal marginal distribution: $q_\theta(a) = p^*(a)$. Note that any CGM that conditions on A [e.g., $q_\theta(X|a)$] can oversample $A = a$ such that the resulting data match the marginal $p^*(a)$.

Misrepresentation bias is due to a misalignment in the conditional distribution of X conditioned on $A = a$; i.e., $p(X|a) \neq p^*(X|a)$. Unlike with underrepresentation bias, where the real data conditional distribution is often replicated, i.e., $q_\theta(X|a) = p(X|a)$, a CGM q_θ that corrects for misrepresentation bias must learn a different conditional distribution from the real data; i.e., $q_\theta(X|a) \neq p(X|a)$.

2.2. The State of the Field: Generative Methods for Tabular Data

Generative models, defined as methods that learn and sample from the full data distribution, have been applied to synthesize realistic datasets across many data modalities. In the last decade (2014–2024), generative models have excelled in the imaging and language fields, leading to the development and proliferation of architectures like variational autoencoders (VAEs) (11), GANs (5), and, more recently, DDPMs (6) and transformers (12). Unfortunately, methods specifically designed for synthesizing tabular data are lacking, and most have resorted to adapting image- or language-specific architectures.

In this section, we first outline existing generative methods (section 2.2.1) and then overview applications toward tabular data (section 2.2.2) and privacy preservation (section 2.2.3).

2.2.1. Overview of generative models. We outline the technical fundamentals of four key generative methods for the unfamiliar reader: GANs, DDPMs, recurrent neural networks (RNNs)(17), and transformers. Two additional generative methods used in our surveyed works, VAEs and Bayes nets, are described in the **Supplemental Appendix**.

2.2.1.1. Generative adversarial networks. One of the original models for deep generative modeling, GANs (5) are optimized via adversarial learning whereby a generative model G learns a

[3]This definition is very similar to the one provided by van Bruegel & van der Schaar (10, p. 2): "a stochastic mathematical model that outputs data and is fitted on real data with the purpose of describing and mimicking (some part of) the real data's distribution."

distribution that approximates the real data distribution p_{real}, as evaluated by learning the discriminator D. The generative model $G(z)$ is a stochastic function of a random latent vector z, typically sampled from the normal distribution; i.e., $z \sim p_Z = N(0, I)$. The loss function is the minimax objective capturing these incentives: $\min_G \max_D \mathbb{E}_{x\sim p_{\text{real}}}[\log D(x)] + \mathbb{E}_{z\sim p_Z}[\log(1 - D(G(z)))]$.

The advantages and challenges of using GANs largely arise from the absence of explicitly estimating the data distribution likelihood. To address these shortcomings, References 13 and 14 reformulate the adversarial objective to increase training stability. Another variant, conditional GAN (15), incorporates a class label as input into the generator and discriminator.

2.2.1.2. Denoising diffusion probabilistic models. DDPMs (also known as score-based generative models) were first introduced by Sohl-Dickstein et al. (16) and popularized by Ho et al. (6) as a powerful method for image generation. Intuitively, DDPMs learn to denoise a noisy perturbation of the original input. The noise perturbation and the denoising estimation are represented by traveling forward and backward, respectively, in a Markov chain of length T. In the forward diffusion process, the continuous-valued input x_0 (e.g., an image) is perturbed over T steps using a known transition equation $q(x_t|x_{t-1})$ that adds Gaussian noise. The final sample is assumed to have a unit normal prior: $x_T \sim p(x_T) = N(0, I)$. In the reverse diffusion process, a model p_θ is trained to denoise the sample by estimating the posterior $p_\theta(x_{t-1}|x_t)$, which is parameterized as a Gaussian distribution. Under this framework, DDPMs can be learned by optimizing the variational bound on the negative log-likelihood:

$$\mathbb{E}\left[-\log p_\theta(x_0)\right] \le \mathbb{E}_q\left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] = \mathbb{E}_q\left[-\log p(x_T) - \sum_{t\ge 1} \log \frac{p_\theta(x_{t-1})}{q(x_{t-1}|x_t)}\right], \qquad 1.$$

where the right-hand side of the inequality can be minimized with respect to θ.

2.2.1.3. Recurrent neural networks. Sequence models, which include RNNs (17) and transformers (12), are a family of methods that process sequential data like language and time series. RNNs are an early form of sequential modeling that iteratively learn and propagate a representation of past sequence information to generate future sequence output. Several variations, including long short-term memory (18), gated recurrent unit (19), and variational (20) RNNs, have since been developed to address shortcomings of the original RNN method.

2.2.1.4. Transformers. In 2017, transformers (12) emerged as a state-of-the-art alternative to RNNs for sequential modeling. Unlike RNNs, which iteratively update a hidden state representation, transformers apply a parallelizable mechanism called self-attention that computes each item's representation based on its relevance to other items in the sequence. Outputs can then be generated by passing these representations through linear layers followed by a decoder (see Reference 21 for a detailed walk-through of transformers). Transformers posses two key advantages over RNNs. First, transformers are efficient to train because representations are calculated in parallel for each item in the sequence. Second, transformers are better able to capture long-term dependencies in the sequence without signal degradation. Despite these advantages, transformers remain computationally intensive with inference time quadratic to sequence length.

2.2.2. Tabular generation. Classic statistical models generated synthetic tabular datasets long before the popularization of neural-network-based approaches. Mixture models (22, 23), sequential decision trees (24), copulas (25), and Bayes nets (26, 27) can be good estimators of simple low-dimensional datasets. Other nonneural-network generative methods include imputation techniques (28, 29) and variations of synthetic minority over-sampling technique (SMOTE) (30), a simple yet effective baseline that generates synthetic points by linearly interpolating real data points. First-generation neural-network-based tabular generative methods like GANs focused on

how to jointly model continuous and categorical variables. Instead of learning the feature space directly, a few early works transformed both variable types into a shared latent space and then learned a GAN over the latent space (31–33).

2.2.3. Privacy-preserving generative methods. To date, a significant motivation to develop generative tabular methods has been to synthesize privacy-preserving sensitive information such as financial, health, or survey data (see Section 2.3.1 for further discussion of data privacy in the medical field). Although quantifying the privacy of an algorithm (or dataset) is challenging and fraught with legal uncertainties (34, 35), several works have proposed generative methods that ensure privacy using theoretical or empirical measures. For example, differential privacy (DP) (36), the most common theoretical metric, bounds the influence each individual data point has on an algorithm's output, thereby limiting the risk of reidentification. References 37–40 all propose DP-guaranteed generative methods for synthetic tabular data.

Federated learning (41, 42) addresses another notion of privacy by ensuring that sensitive data remain in-house during training. For instance, in the realistic scenario where several clients (e.g., hospitals) each hold exclusive access to sensitive health data, federated learning aggregates each client's model updates to train a single generative method. References 43–45 develop GANs that use federated learning to securely aggregate updates over multiple sensitive tabular datasets.

2.3. The State of the Field: Medical Applications

The primary motivation of synthetic medical data thus far has been to preserve patient privacy, which we summarize in Section 2.3.1. However, the potential of generative models to improve medical data has not been fully explored (10, 46), and in Section 2.3.2 we highlight emerging applications toward data debiasing and synthesizing digital twins.

2.3.1. Private and reproducible medical research. Due to the highly sensitive nature of healthcare data and stringent regulations like the Health Insurance Portability and Accountability Act (HIPAA), many sources of valuable medical information are inaccessible. Synthetic and deidentified data would thus democratize access to medical data and enable applications such as secondary analyses of clinical trial data (47), regulating medical devices (48), investigating urgent pandemics like COVID-19 (49), testing medical software (50), training medical students on realistic patient data (51), and many others as highlighted in References 52–55.

The availability of synthetic data also encourages research reproducibility and fosters collaboration (56), as evidenced by the proliferation of publicly available software tools and synthetic datasets. For instance, two popular software tools, MDClone and Synthea (57), have been used to synthesize datasets in several medical research studies (58).

2.3.2. Debiasing medical datasets. Real-world medical datasets often misrepresent historically marginalized populations and lack adequate sample size for rare health conditions. These biases disproportionately harm certain populations, which can lead to incorrect conclusions about treatment efficacy or the censorship of entire groups (2, 59, 60). For example, the exclusion of pregnant women, children, and marginalized racial groups from clinical trials has resulted in adverse or ineffective drug responses (3, 61, 62) such as the reduced efficacy of warfarin for Black and Hispanic patients (63). The need for equitable and representative data has prompted US Food and Drug Administration initiatives to recruit diverse clinical trial cohorts (64) and the creation of the All of Us biobank. Nevertheless, biased medical data continue to be a pervasive issue.

One promising application of generative models is to correct for these harmful biases. Although there are many forms of data bias (65), most works have focused on underrepresentation

bias (defined in Section 2.1.2) by augmenting imbalanced medical labels with synthetic data generated from CGMs (66–69). While these approaches often replicate the real data distribution, a few works have sought to improve the real data by removing other forms of data bias. Generative approaches for active data debiasing include weighted sampling during training (70–72), transforming the sampling distribution post hoc (73), intervening on a causal model of the data (74, 75), and enforcing algorithmic fairness (76, 77). Applying these debiasing approaches to generate more representative data is one step toward ensuring that medical research serves those who have been historically neglected.

2.3.3. Simulating the unobserved: digital health twins. Another emerging and compelling medical application of CGMs is digital health twins, which virtually simulate a patient's physiological state, including their response to hypothetical interventions. Indeed, several recent works have highlighted the promise of generative methods, specifically CGMs, to simulate digital health twins by conditioning on patient-specific attributes (46, 78–81).

By dynamically integrating multimodal data sources, digital health twins have the potential to advance precision medicine and drug development (82, 83). For instance, randomized controlled trials, the gold standard for validating treatment efficacy in drug development, are expensive and can raise difficult ethical questions (84). On the other hand, CGMs offer the affordable and potentially more ethical option of simulating virtual patient cohorts for in silico trials (46, 79, 83, 85). CGM-simulated digital health twins could also help predict disease progression or forecast patient trajectories (78, 79, 86, 87). Finally, generated counterfactual patient data (88, 89) (as defined in Section 4.1.2.3) can model the effect of a hypothetical intervention, thereby aiding in treatment optimization and analyzing the effect of attributes like race on a patient's quality of care (79, 85, 87).

2.4. Related Works

In **Table 1**, we list 16 related review papers that address generative methods for tabular data and assess the following criteria: (*a*) Was the review focused on the medical domain? (*b*) Did the review highlight the difficulties specific to modeling tabular data? (*c*) Were methods for tabular time-series generation included? (*d*) Were modern, state-of-the-art ML methods beyond autoencoders and GANs covered? and (*e*) Did the review cover applications outside of privacy preservation, such as data enhancement?

We note a few trends from the reviews. As expressed in Reference 53, there are few methods for generating longitudinal medical data, despite the data type's ubiquity. Furthermore, the vast majority of prior reviews (all but Reference 87) do not cover modern generative methods beyond classical statistical methods, GANs, and autoencoders. Finally, while most of the works highlighted privacy preservation applications, a few reviews recognized the potential of generative methods to improve data. References 10, 69, and 90 discuss generative methods that enforce fairness and augment for underrepresented groups, and References 46, 78, and 79 highlight the potential of simulating digital health twins. Most similar to our work, Reference 85 reviews medical applications outside of privacy but limits the coverage to GANs.

In this review, we contribute the following:

1. We highlight the importance of CGMs for tackling two applications relevant to precision medicine: debiasing datasets and simulating digital health twins.
2. We review models that generate either static or longitudinal (time-series) tabular data, and we discuss how these methods handle the technical difficulties specific to this data modality.
3. We cover emerging generative methods, including DDPMs and transformers.

Table 1 Related review papers

Reference	Medical[a]	Tabular[b]	Time series[c]	Modern models[d]	Improve data[e]
90			×		×
85	×	×	×		×
91		×			
92	×	×			
10					×
78	×				×
93	×	×	×		×
53	×	×	×		
52	×				×
46	×				×
87	×			×	×
79	×				×
94	×	×	×		
69		×			×
95	×	×	×		
96		×			

[a]Was the review focused on the medical domain?
[b]Did the review highlight the difficulties specific to modeling tabular data?
[c]Were methods for tabular time-series generation included?
[d]Were modern, state-of-the-art machine learning methods beyond autoencoders and generative adversarial networks covered?
[e]Did the review cover applications outside of privacy preservation such as data enhancement?

3. SELECTION

3.1. Scope

We next define the inclusion/exclusion criteria of the surveyed methods. The three criteria for inclusion are time, topic, and data.

1. Time: The article was published between January 2019 and May 2024.
2. Topic: The article must propose a new method that fits Definitions 1 and 2 of a CGM.
3. Data: We allow for the data to be tabular (a mix of continuous and categorical variables), completely categorical (e.g., binary medical codes), or completely continuous (e.g., biomarkers as a time series).

The four criteria for exclusion are model intention, language, availability, and prior knowledge.

1. Model intention: As per Definition 2, we exclude CGMs solely intended for prediction instead of synthetic data generation.
2. Language: The article must be in English.
3. Availability: The article must be publicly available.
4. Prior knowledge: As classified in Reference 53, the method must be purely data driven, whereby only the training data are used for learning the generative model, and there is no integration of domain knowledge or human expertise.

To incorporate a variety of potentially useful methods, we did not require an article to be motivated by medical applications. Additionally, while peer-reviewed articles were preferred, we

make a few exceptions and use our judgment to assess article quality. Although requiring an article to be peer-reviewed ensures no false positives, it can certainly lead to false negatives. We note in **Table 2** which articles were not peer reviewed at the time of our search.

3.2. Search Details

We ran our search first in November 2023 and again in May 2024 using Google Scholar, PubMed, arXiv, and Science Direct as sources. The specific search terms are listed in the **Supplemental Appendix**. After screening titles and abstracts, we collected a list of approximately 100 articles that fit the inclusion/exclusion criteria. We then read and narrowed down the list to the final 43 papers.

4. ANALYSIS

The 43 selected methods are outlined in **Table 2**. In Section 4.1, we discuss how these works address two emerging applications toward precision medicine: data debiasing and digital health twins. In Section 4.2, we review how these methods tackled the technical issues of tabular health data. We conclude in Section 4.3 with a brief overview of how CGMs are evaluated.

4.1. CGMs Defined by Use Case

We next analyze the relevant 43 CGM methods based on their relevancy to two medical applications: data debiasing and digital health twins.

4.1.1. Data debiasing. In this section we highlight CGM approaches that correct for (*a*) underrepresentation bias and (*b*) misrepresentation bias (both of which are defined in Section 2.1.2).

4.1.1.1. Balancing group underrepresentation through data augmentation. Although any CGM that conditions on the underrepresented group could correct for underrepresentation bias via data augmentation, here we cover works that explicitly evaluate the effect of augmentation on downstream prediction tasks. With respect to the underrepresented group, only two works consider a historically marginalized group, while the rest address label imbalance (e.g., a rare disease). HealthGen (102) demonstrates that upsampling underrepresented groups, such as those with government insurance or who identify as Hispanic, can occasionally improve performance in clinical prediction tasks. Another method, MTGAN (99), generates time-series data for rare diseases and shows that pretraining a classifier on their synthetic data outperforms other generative models. Reference 101, which discusses RelDDPM, and Reference 98 both propose and evaluate DDPM-based CGMs that upsample multiple conditional attributes. Despite the promise of CGM-based data augmentation, performance gains are currently inconsistent, and SMOTE (30), a simple baseline based on point interpolation, remains a strong competitor (98, 100, 101).

4.1.1.2. Correcting for misrepresentation bias in real data. Other methods address data misrepresentation by learning a debiased version of the original data distribution, often by transforming the feature distribution or enforcing algorithmic fairness. As an example of the former, CTGAN (106) upweights the importance of underrepresented groups by sampling points during training based on the logarithm of the group frequency, a correction akin to oversampling or reweighting. An ablation study shows that this approach, called training-by-sampling, increases the overall F1 score by almost 18% when compared to sampling the true group frequency. Similarly, CTAB-GAN+ (108) applies a logarithmic transform to prioritize learning a feature's long tails, which often contain outlier points.

BD08_Art02_Liu ARjats.cls December 18, 2024 11:55

Table 2 Surveyed conditional generative model works

Paper	Medical application[a]	Data type[b]	Generative model[c]	Variable types modeled[d]	Tabular modeling[e]	Missingness[f]
MaskEHR (97)	DHT–prospective, underrepresentation bias	Regular time series	GAN, RNN	Categorical	NA	Omission
Reference 98	Underrepresentation bias	Static	DDPM	Categorical, continuous	Embedding	None
MTGAN (99)	Underrepresentation bias	Regular time series	GAN, RNN	Categorical	NA	None
Reference 100	Underrepresentation bias	Static	GAN	Categorical, continuous	Embedding	None
RelDDPM (101)	Underrepresentation bias	Static	DDPM	Categorical, continuous	Categorical binary encoded	None
HealthGen (102)	Underrepresentation bias	Regular time series	AE, RNN	Continuous	NA	Binary variables
Bt-GAN (103)	Misrepresentation bias	Static	GAN	Categorical	NA	None
FLMD (104)	Misrepresentation bias, underrepresentation bias	Regular time series	AE, RNN	Categorical, continuous	Continuous binned, categorically binary encoded	None
HALO (110)	Underrepresentation bias, DHT–prospective	Irregular time series	Transformer	Categorical	NA	None
GOGGLE (105)	DHT–CF, misrepresentation bias, underrepresentation bias	Static	AE, PGM	Categorical, continuous	Explicit	None
CTGAN (106)	Misrepresentation bias	Static	AE, GAN	Categorical, continuous	Mode-specific normalization	None
Reference 107	Misrepresentation bias	Static	GAN	Continuous	NA	None
CTAB-GAN+ (108)	Misrepresentation bias	Static	GAN	Categorical, continuous	Mode-specific normalization	Binary variables
DPTVAE (109)	Misrepresentation bias	Static	AE	Categorical, continuous	Mode-specific normalization	None
Reference 111	DHT–prospective	Regular time series	AE, RNN	Continuous	NA	None
POPCORN (112)	DHT–prospective	Irregular time series	RNN	Categorical	NA	None
TimeGrad (113)	DHT–prospective	Regular time series	DDPM, RNN	Continuous	NA	None
EVA (114)	DHT–prospective	Regular time series	AE	Categorical	NA	None
StoCast (115)	DHT–prospective	Irregular time series	AE, RNN	Categorical, continuous	Continuous binned	None
TWIN (116)	DHT–CF, DHT–prospective	Regular time series	AE	Categorical	NA	None
Reference 117[g]	DHT–prospective	Regular time series	None	Categorical, continuous	Categorical binary encoded	None
EHR-M-GAN (118)	DHT–prospective	Regular time series	AE, GAN, RNN	Categorical, continuous	Embedding	None
VHP (119)	DHT–prospective	Irregular time series	AE, GAN, RNN	Continuous	NA	None
SCouT (120)	DHT–CF, DHT–prospective	Regular time series	Transformer	Categorical, continuous	Categorical binary encoded	None
SynTEG (121)	DHT–prospective	Regular time series	GAN, RNN	Categorical	NA	None

(*Continued*)

Table 2 (*Continued*)

Paper	Medical application[a]	Data type[b]	Generative model[c]	Variable types modeled[d]	Tabular modeling[e]	Missingness[f]
Clinical-GAN (122)	DHT–prospective	Regular time series	Transformer, GAN	Categorical	NA	Omission
CEHR-GPT[g] (123)	DHT–prospective	Irregular time series	Transformer	Categorical	NA	Omission
Reference 124	DHT–prospective	Regular time series	RNN	Categorical	NA	None
RNN-ODE (125)	DHT–prospective	Irregular time series	AE, RNN	Categorical, continuous	Embedding	None
PromptEHR (126)	DHT–prospective	Regular time series	Transformer	Categorical	NA	Omission
Reference 127	DHT–prospective	Irregular time series	RNN	Categorical, continuous	Continuous binned	Omission
VCNET (128)	DHT–CF	Static	AE	Categorical, continuous	Categorical binary encoded	None
Reference 129[g]	DHT–CF	Static	AE	Categorical, continuous	Categorical binary encoded	None
CeFlow (130)	DHT–CF	Static	None	Categorical, continuous	Embedding	None
CounteRGAN (131)	DHT–CF	Static	GAN	Categorical, continuous	Categorical binary encoded	None
SMOOTH-GAN (132)	DHT–CF	Static	GAN	Categorical, continuous	Categorical binary encoded	None
LongGAN (133)	None	Regular time series	AE, GAN, RNN	Continuous	NA	None
C3-TGAN[g] (134)	None	Static	GAN, PGM	Categorical, continuous	Mode-specific normalization	None
TabDDPM (135)	None	Static	DDPM	Categorical, continuous	Embedding	None
REaLTabFormer[g] (136)	None	Static	Transformer	Categorical, continuous	Continuous binned	Omission
GReaT (137)	None	Static	Transformer	Categorical, continuous	Continuous binned	Binary variables
TabMT (138)	None	Static	Transformer	Categorical, continuous	Embedding	Binary variables
Synthsonic (139)	None	Static	PGM	Categorical, continuous	Explicit	None

Abbreviations: AE, autoencoder; CF, counterfactual generation; DDPM, denoising diffusion probabilistic model; DHT, digital health twins; GAN, generative adversarial network; NA, not applicable; PGM, probabilistic graphical model; RNN, recurrent neural network.

[a]None indicates none of the above. Prospective represents future health trajectory generation.

[b]See Section 2.1.1 for definitions of irregular and regular time series.

[c]None indicates none of the above.

[d]Categorical includes only binary variables.

[e]If applicable (i.e., both categorical and continuous variable types considered), how tabular data were modeled. Continuous binned indicates continuous variables are binned, making the data fully binary; categorial binary encoded indicates categorical variables are 1-hot encoded [or, in one case, analog-bit encoded (101)] and continuous variables are typically scaled to a [0, 1] range; embedding indicates separate embedding heads, either learned or fixed, for continuous and categorical variables; explicit indicates continuous and categorical variables separately modeled, typically with PGMs; and mode-specific normalization indicates the method introduced in Reference 106.

[f]How paper handles missing-not-at-random missingness. Binary variables indicates missingness is modeled as an extra binary variable for each variable; omission indicates missingness is implicit by omission.

[g]Not peer-reviewed.

Another approach to learning a debiased data distribution is to enforce a specific algorithmic fairness metric via an auxiliary loss function. For example, Bt-GAN (103) enforces statistical parity (140) by penalizing the mutual information between protected group status and the remaining generated features. FLMD (104) applies a two-stage framework that controls for unobserved confounders using the deconfounder method (141) and then optimizes for counterfactual fairness (142) under a hypothetically different demographic attribute. Similarly, Reference 107 proposes a GAN-based method that generates a counterfactually fair dataset. Requiring only a preliminary causal model of the data, Reference 107 shows the resulting synthetic dataset can train a classifier that upholds fairness and accuracy on the original biased data.

4.1.1.3. Potential and cautionary recommendations for development. While a promising application, data debiasing using CGMs requires determining which biases are unwanted. For example, Bt-GAN, FLMD, and similar fairness-enforcing generative methods (76, 77) assume a specific algorithmic fairness metric and prediction task to debias and are thus not generalizable to other metrics or tasks. An additional challenge is the lack of standardized evaluation, exemplified by the discordance of metrics and baselines for benchmarking augmentation efficacy.

4.1.2. Digital health twins. A second medical application of CGMs is digital health twins. In this section, we highlight approaches that use CGMs for either (*a*) prospective generation of patient trajectories or (*b*) counterfactual generation by simulating patient data under a hypothetical intervention.

4.1.2.1. Prospective patient generation: methods. By conditioning on past medical history, CGMs can forecast a patient's health trajectory by autoregressively generating all variables over several timesteps. We further split the identified works into those that model regular versus irregular time-series data.

First, as defined in Section 2.1.1, regularly sampled time series contain all K variables at discrete, evenly spaced steps. Regular time series are less common in real-world health data, and most of the surveyed works transform irregular into regular time series by aggregating each variable over evenly spaced time intervals.

Although most of the surveyed works (97, 102, 104, 111, 118, 121, 124) model regular time series with RNNs (17), only three (120, 122, 126) use transformers. Because transformers were developed to conditionally predict the next discrete variable for language generation, application to tabular longitudinal health data requires preprocessing the data into the correct format. PromptEHR (126) massages EHR data into visit-level units to feed into the encoder–decoder transformer BART (143). Using a GAN framework, Clinical-GAN (122) generates longitudinal trajectories from a transformer-based generator. However, both approaches are constrained to discrete medical codes.

Many works further optimize synthetic time-series realism by adding an auxiliary discriminator loss, thereby forming a GAN structure. Clinical-GAN, EHR-M-GAN (118), MaskEHR (97), and SynTEG (121) all train a discriminator to detect real from fake trajectories generated by a sequential model, where only Clinical-GAN uses a transformer as the generator.

A more realistic form of medical data is irregularly sampled time-series data. StoCAST (115), HALO (110), CEHR-GPT (123), and the method described in Reference 127 all model irregularly sampled data as inherently regularly sampled by utilizing sequence models like RNNs or transformers. However, to incorporate continuous-valued timestamps of each medical observation, three of these methods (110, 115, 127) add sojourn time (the time between visits) as an additional continuous variable to model. Alternatively, CEHR-GPT represents time between visits by inserting the appropriate time-specific tokens.

However, merely incorporating a continuous-valued timestamp variable may not robustly model the distribution of irregularly sampled data, as the learned data density function is still discretely sampled. As a potential solution, POPCORN (112), VHP (119), and RNN-ODE (125) incorporate temporal point processes (144) to explicitly model the density of variable occurrence over a continuous function of time. However, neither POPCORN nor VHP handles both continuous and categorical variables, and RNN-ODE only models continuous regularly sampled and categorical irregularly sampled data.

4.1.2.2. Prospective patient generation: potential and cautionary recommendations for development. Approaches like those described in References 115, 123, and 126 that condition on both patient history and immutable demographic information highlight the potential of CGMs for simulating personalized trajectories. However, applying CGMs to tabular time-series data remains difficult. For one, only a few of the surveyed approaches (104, 115, 117, 118, 120, 125, 127) model both continuous and categorical time-series data, a challenge further discussed in Section 4.2.1. Second, although many methods optimize for generating realistic longitudinal datasets (97, 110, 114, 116–118, 121, 123, 127), often by incorporating GANs, the field of time-series modeling prioritizes localized forecasting. This problem might be addressed by leveraging transformers, which are adept at capturing long-term dependencies. Finally, modeling irregular time-series data remains an unsolved problem.

4.1.2.3. Retrospective patient generation: methods. Intuitively, counterfactual refers to a hypothetical scenario had the observed data been generated under different conditions, usually by changing a single variable (145). More concretely, counterfactual generation requires first intervening on this variable (which we refer to as treatment or label intervention), forcing it to assume a specific value in the causal model, contrary to what was observed.[4] Robust estimation of the counterfactual under label intervention enables treatment effect estimation, which is useful for a broad range of medical applications. However, generating counterfactuals is challenging, as it relies on several strong assumptions and requires a causal model of the data (145).[5]

One method, GOGGLE (105), estimates the causal model via a message-passing algorithm and thus could reasonably be a causally accurate counterfactual generator. However, this application was not empirically assessed and furthermore requires learning a separate GAN for each conditional generation task.

Sacrificing rigor for tractability, many approaches forgo the need for a causal model and approximate the process of label intervention. References 128 and 130–132 frame the intervened label as a conditional input into a CGM. An observed data point is then transformed into the approximate counterfactual by altering only the label in the conditional input to be different than what was observed.

Unlike static data, time-series data contain a natural notion of causality based on temporal occurrence. Even if the true causal model is not known, CGMs approximating counterfactual generators can condition on both patient history and the desired medical intervention to generate future hypothetical data. SCouT (120) uses a transformer to simulate a patient's future trajectory as if they had hypothetically received the control treatment. Given a target patient and the desired

[4]As pointed out by Abroshan et al. (107) and Pearl's three rungs of causality (145), label intervention is not always equivalent to counterfactual generation. However, we consider this distinction out of scope for this review and refer to them interchangeably.

[5]Note that generating counterfactually fair data (142) is roughly the opposite of generating counterfactuals. Briefly, counterfactual data generation captures the effect of the intervened label on the rest of the data, whereas counterfactually fair data remove the impact of the intervened label.

hypothetical treatment at a specific timestep, TWIN (116) searches the observed data for the top-k patients who are most similar to the target patient but received the desired treatment. An autoencoder then takes as input the target's history, the top-k most similar patients, and the desired treatment to generate the future counterfactual data for the target patient. Although promising counterfactual approximators, both SCouT and TWIN require sufficiently diverse patient data.

4.1.2.4. Retrospective counterfactual patient generation: potential and cautionary recommendations for development. Counterfactual-generating CGMs could enable personalized and equitable medicine. For instance, the approaches discussed in References 116 and 120 are promising in silico simulators for synthetic control arms, and those in References 128 and 130–132 could help interpret how race affects a patient's clinical outcomes. However, there is a necessary trade-off between causal precision and efficiency. Approaches that approximate label intervention (128, 130–132) are computationally efficient but lack causal rigor. Other challenges to counterfactual CGMs include the difficulty of modeling nontrivial interventions, assumed access to a label classifier, and potential counterfactual implausibility. For example, approximating counterfactual generation could lead to impossible outcomes like a patient whose age decreases after a treatment. The concept of counterfactual plausibility is considered by References 131 and 134 and could be further addressed by constraint-based generative methods (146).

4.2. Conditional Generative Models Defined by Technical Capacity

The practical use of CGMs requires the ability handle real-world tabular medical data. Unfortunately, aside from probabilistic graphical models like Bayes nets, which are computationally feasible for only a modest number of variables, there is a dearth of generative methods that jointly learn continuous and categorical data. Until an architecture is developed specifically for tabular data, generative methods must compromise between distribution accuracy and computational efficiency.

Tabular medical data pose several modeling challenges, as they often involve (*a*) multiple variable types, (*b*) missing-not-at-random (MNAR) features, (*c*) high-dimensional data, (*d*) small sample sizes, (*e*) time-series data (as covered in Section 4.1.2), (*f*) imbalanced or biased distributions (as covered in Section 4.1.1), and (*g*) a high number of unobserved confounders. We first highlight how the surveyed works handle issues *a–d* and then discuss the trade-offs of different CGM architectures.

4.2.1. Handling mixed-type tabular data. Twenty-five of the surveyed works model both continuous and categorical data. To handle mixed-type data as model input, many works concatenate one-hot encoded categorical data with continuous data that are either discretized into bins (115, 127) or scaled to a [0, 1] range (104, 117, 128, 129, 131, 132). While such preprocessing is simple and enables learning a single model, merging two distinct variable types into the same feature space sacrifices information, particularly for continuous variables. A slightly improved version of continuous data discretion is mode-specific normalization (106, 108, 109), which models each continuous variable as a probabilistic representation of a fitted variational Gaussian mixture.

Modeling mixed-type data using transformers, which were developed for discretized data like language, poses another challenge. References 120, 136, and 137 discretize each continuous value into its own token to be embedded alongside categorical variables. Alternatively, TabMT (138) proposes an embedding method where continuous variables are first quantized via k-means and then transformed into a numerically friendly ordered embedding.

References 100, 118, 125, 130, and 135 model continuous and categorical data by separate, potentially learnable, embedding functions that project each variable type into a shared latent space.

EHR-M-GAN (118) trains separate autoencoders for categorical and continuous data and adds an auxiliary loss to enforce latent space similarity. TabDDPM (135) and the method discussed in Reference 98 map continuous data using Gaussian diffusion and categorical data using multinomial diffusion (147) to a latent space with a unit normal prior. While these approaches preserve the underlying distributions of each variable type, they independently learn the latent information of categorical and continuous variables even though the variables are likely not independent of each other.

Finally, the methods discussed in References 105 and 139 both learn a probabilistic graphical model (PGM) over the data, which naturally handles mixed-type data by explicitly learning each variable's conditional distribution.

4.2.2. Handling missing-not-at-random features. One challenge specific to tabular healthcare datasets is the presence of MNAR features. Most works assume unobserved features are missing completely at random and thus can be dropped or imputed. However, several studies have refuted this assumption by proving missingness can be informative of measured clinical outcomes (102, 148, 149), e.g., reflecting a patient's access to healthcare (150). Thus, handling real-world medical datasets requires a generative model that learns structural patterns of missingness.

Of the surveyed works, References 97, 102, 108, 122, 123, 126, 127, 136, 137, and 138 preserve missingness in the learned distribution. Doing so is relatively straightforward for categorical variables, where missingness can be appended as a potential class for every variable. For example, the transformer-based method GReaT (137) incorporates a missingness token into its discrete-valued vocabulary. Missingness can also be handled by omission in the case of autoregressive CGMs that output data without a fixed order. ClinicalGAN (122), MaskEHR (97), and the method discussed in Reference 127 output any number of the total K variables at each visit, where omitted variables are implicitly missing.

Representing MNAR continuous features is less straightforward. HealthGen (102) explicitly models a missingness mask for each patient and uses a masked loss over observed data to avoid representing missing continuous variables. CTAB-GAN+ (108), inspired by mode-specific normalization (106), appends a binary missingness variable as a potential mode when representing continuous variables.

4.2.3. High-dimensional data. Although real-world medical datasets like EHRs typically contain a large number of features, only four of the surveyed works (104, 105, 125, 132) generate high-dimensional tabular data (defined as $K > 100$).[6] SMOOTH-GAN (132) learns 166 static variables from the Cerner HealthFacts Database, and RNN-ODE (125) and FLMD (104) both learn tabular time-series data from MIMIC-III with 104 and 941 variables per timestep, respectively.

Other methods considering high-dimensional data are limited to binary variables like disease (e.g., ICD-10) or prescription codes. For instance, transformer-based CGMs like those described in References 97, 110, and 122 encode thousands of discrete medical codes into a fixed or learnable embedding space.

4.2.4. Small datasets. Although some medical datasets, e.g., EHRs, have a large sample size, datasets collected from clinical trials or disease-specific cohorts are often small due to monetary cost, patient risk, or phenotype rarity. Applying CGMs for rare disease augmentation or synthetic control arms thus requires the ability to learn from but not overfit to small datasets.

[6]We consider a time-series dataset with T steps and M features has $K = M$ features, not $K = T \cdot M$.

Approximately 40% of the surveyed works evaluate a dataset with <1,000 patient data points. The transformer-based method GReaT (137) reduces overfitting by using a pretrained large language model on a small static dataset of 8 features and 954 samples. TWIN (116) considers data with two variables from only 77 patients. To prevent overfitting, TWIN learns a weighted aggregation of a patient with their most similar data points, which feeds into a small generative model. Finally, PGMs like GOGGLE (105) are ideal for learning small datasets, as they regularize variable dependencies and thus require minimal parameters.

4.2.5. Generative models: pros and cons. In this section, we categorize the surveyed works by the generative methods used and discuss each method's pros and cons.

4.2.5.1. Generative adversarial networks. Fourteen of the surveyed works incorporate GANs into their generative framework. GANs have the benefit of intuitive appeal, historical precedent, and flexibility in selecting the underlying model for the discriminator and generator. Additionally, by optimizing over an adversarial loss and thus avoiding explicit likelihood approximation, GANs offer greater flexibility in the learned distribution and are typically much faster to sample from compared to PGMs or DDPMs. As shown by the surveyed works, GANs are able to generate time-series data by learning an autoregressive generator to simulate full or partial trajectories and an auxiliary discriminator loss to ensure trajectory realism (97, 99, 118, 119, 133).

However, GANs have limitations. The absence of explicit likelihood estimation complicates performance assessment, and the adversarial optimization hinders model convergence. Despite attempts to stabilize performance (13, 14), GANs remain infamously difficult to train and often suffer from mode collapse, whereby only a limited variety of outputs are generated. Furthermore, GANs are constrained to only handle continuous data, as it is difficult to backpropagate gradients from the discriminator to a generator that outputs discrete values. To apply GANs to categorical or tabular data, a few of the surveyed works pass the probabilistic generator output into the discriminator, which now has the easy task of separating real discrete from synthetic continuous data (131, 132). Other more robust solutions are to learn the GAN over a continuous latent space from embedded categorical data (100, 118, 121) or to apply the Gumbel-Softmax trick (151).

4.2.5.2. Denoising diffusion probabilistic models. DDPMs are an emerging and promising generative method used by four of the surveyed works (98, 101, 113, 135). To handle tabular data, TabDDPM (135) and the method described in Reference 98 leverage multinomial diffusion (147). In a follow-up work applying TabDDPM to tabular EHR data, Ceritli et al. (152) demonstrated that DDPMs outperformed simple GAN baselines in terms of data utility and realism, although they fall short on privacy metrics. DDPMs could also be used to model time-series data. For instance, although limited to continuous variables, TimeGrad (113) trains a DDPM to predict next timestep data conditioned on prior information. By optimizing an explicit likelihood estimation instead of adversarial loss, DDPMs outperform GANs in imaging in terms of sample diversity and avoiding mode collapse (153, 154). Although a similar diversity assessment for tabular datasets is currently lacking, DDPMs present a promising alternative to GANs that could better represent rare diseases and marginalized groups.

One challenge, however, in applying DDPMs to the tabular domain is the assumption of continuous-valued data when performing Gaussian diffusion, which perturbs the data by adding and removing Gaussian noise. Although variants of categorical-valued diffusion (147, 155) exist, categorical synthetic data have not undergone the same rigorous validation as continuous data have in imaging. Another disadvantage of DDPMs is their computational intensity, particularly during inference. Combined with the requirement to match the latent space dimension to the input size [i.e., $\dim(x_0) = \dim(x_i)\ \forall i = 1, \ldots T$], which might be prohibitively large for data like EHRs, DDPMs are among the slower methods for generating synthetic data.

4.2.5.3. Recurrent neural networks and transformers. The vast majority of the surveyed works modeling time-series data use RNN-based methods. StoCast (115) learns an RNN to capture patient history and then generates next timestep data by conditioning on the RNN hidden state and patient-specific fixed variables. Only 5 of the 20 surveyed works for time-series generation employ transformers (110, 120, 122, 123, 126).

Transformers can also be applied to generate static tabular data. REaLTabFormer (136) fixes an arbitrary variable order and uses the decoder-only transformer GPT-2 (156) to autoregressively generate relational and nonrelational static tabular data. TabMT (138) optimizes the bidirectional encoder-only transformer BERT (157) using masked language modeling, where conditional sampling occurs by predicting all the masked variables.

Transformer-based CGMs exhibit many advantages. First, transformers are able to model high-dimensional data by tokenizing and aggregating inputs. Second, they are superior to RNNs at capturing long-term variable dependencies, which is advantageous for modeling longitudinal EHR data where past health conditions influence the current health state. Third, transformers can easily represent MNAR features.

However, it is unclear how good transformers are at understanding and generating continuous variables. Additionally, transformers require sufficient computational resources and dataset size for training. When evaluating run time, Gulati & Roysdon (138) showed their transformer-based approach took 144 times longer to train and 66 times longer for inference compared to VAE and GAN baselines. For small medical time-series datasets like clinical trials, RNNs might be a more appropriate model.

Recently, there has been a trend toward transformer-based foundation models, whereby a large model is first pretrained over massive but typically closed-source datasets and then fine-tuned for a specific application, like data generation (137). While foundation models can potentially inherit prior medical knowledge learned during pretraining, the lack of quality and transparency in the pretraining dataset risks introducing false information and harmful biases into the data generation process.

4.3. Evaluation

Evaluating generative methods remains challenging. Unlike supervised models, which have a well-defined objective, generative models are a tool whose product—synthetic data—should be amenable to a range of analyses unlikely to be known a priori. If the intended use case of the model is replicating the real data, an intuitive measure of success is the distance between the real and learned joint data distributions via a likelihood function. Conversely, if the goal is data enhancement, as with data debiasing, success is achieved by simultaneously minimizing the likelihood function and achieving the enhancement objective. Unfortunately, computing likelihood is often intractable and unlikely to be comprehensive.

The ambiguity of synthetic data has led to a proliferation of metrics for evaluating the quality of generative models and synthetic data. Alaa et al. (158) proposed an evaluation framework based on fidelity, generalizability, and diversity, whereas other works suggested utility, resemblance, and privacy (93, 159). Similarly to Vallevik et al. (160), we identify five categories for evaluating synthetic tabular data: (*a*) utility, (*b*) resemblance, (*c*) privacy, (*d*) diversity, and (*e*) computational cost.

4.3.1. Evaluation categories. We refer the reader to References 25, 53, 85, 92, 93, 159–163 for more thorough discussions on synthetic data evaluation and metrics.

4.3.1.1. Utility. Utility aims to measure synthetic dataset performance for any downstream task that the real data would be used for. In practice, evaluating utility requires training an ML prediction model and comparing performance across real and synthetic data. One test might

ensure an ML classifier trained on real data can generalize to synthetic data; other variations of this train-test framework are possible and covered in References 93 and 159. If the model is intended to resemble the real data, the optimal utility is for synthetic data to perform comparably to real. However, if the objective is to debias or augment the real dataset, it is preferable for the synthetic dataset to perform better.

4.3.1.2. Resemblance. Resemblance metrics evaluate how statistically similar the synthetic data are to the real data, such as by testing univariate or multivariate distribution similarity. Like utility, the value of resemblance depends on the model's intended purpose. If the model is trying to mitigate harmful biases, requiring comparable resemblance to real data might not be appropriate.

4.3.1.3. Privacy. Although not the focus of this review, it may be crucial to ensure the synthetic data protect sensitive information of real individuals. Privacy metrics measure this disclosure risk and are explored further in References 52, 53, 85, and 93.

4.3.1.4. Diversity. One category of evaluation metrics that necessitates more attention is diversity, a term we use to also encompass fairness. A few works (158, 164) measure diversity in terms of sample coverage, where it is ideal for the real data to be sufficiently covered (in feature or latent space) by synthetic data. On a similar thread, References 10, 160, 162, 165, and 166 highlight the importance of fairness metrics in ensuring identified marginalized groups have the same quantity and quality in both synthetic and real data. For example, Bhanot et al. (166) proposed a log disparity score to measure marginalized group representation in both static and longitudinal healthcare data. Other works (74, 103, 104) assume that adherence to specific algorithmic fairness metrics, like statistical parity, indicates a successful synthetic dataset.

4.3.1.5. Computational cost. Hernandez et al. (93) and Vallevik et al. (160) identified the importance of evaluating the computational cost and carbon footprint of generative models. Computational complexity metrics—which include listing the number of model parameters, training and inference time, and computational resources used—are helpful for both practical implementation and tracking a model's environmental impact. For instance, transformer-based methods are a promising approach for generating high-quality tabular data, but they require a large number of computational resources and thus might not be optimal for routine use in low-resource settings.

4.3.2. Recommendations for development. Given the wide range of medical applications, we recommend further investigation into developing health-specific evaluation metrics for CGMs. For instance, a generative model intended to preserve patient privacy should have different evaluation criteria than one synthesizing digital health twins. In the latter scenario, the priority of evaluation might be the detection of rare and lethal health outcomes over privacy or computational complexity. Thus, developing CGM evaluation methods in each medical setting requires a collaborative effort among stakeholders to determine what constitutes a success.

We also highlight the importance of interpretable evaluation, especially for end users like clinicians who might have limited understanding of metric implication. Metric performance often involves trade-offs (e.g., fairness may be inversely related to utility), complicating the ability to claim one model is better than another. Therefore, it is crucial to be able to interpret and translate evaluation claims into practical advice. We discuss this and other limitations below.

5. LIMITATIONS

Despite their potential, CGMs carry risks that could exacerbate existing problems for the medical community, and it is critical to proceed with research that is cautious, thoughtful, and collaborative. Here, we address the limitations of current CGMs, concerns for the field, and areas for further research. We categorize these into technical, medical, and ethical limitations.

5.1. Technical

We first highlight the need and technical limitations of current CGMs in modeling missing features and causal effects.

5.1.1. Modeling missing-not-at-random features. As highlighted in Section 4.2.2, generative models unable to represent MNAR features risk losing essential information and could further marginalize groups typically associated with these features (102, 148–150). One recommended area of technical development is appropriate modeling of MNAR features, an issue also explored in References 167–169. We also recommend thorough analysis of the missingness mechanism before assuming missing values can be imputed or removed.

5.1.2. Causally aware generative models. In many settings that require decision-making, it is essential to guarantee or interpret causal effects. However, most CGMs only capture correlations between variables and thus conditioning on a label do not guarantee any meaningful causal effect. Alternatively, PGMs like Bayes nets capture causal relationships, but existing methods only approximate the causal model and are limited to a modest number of variables (105, 139). Developing causally aware generative models, as recently explored by Wen et al. (170) and Rahman et al. (171) and in the review by Komanduri et al. (89), would promote the application of synthetic data for medical research, including clinical decision-making.

5.2. Medical

Generating synthetic patient health data requires developing targeted CGMs for medical applications. Here, we emphasize the importance of multimodal modeling and integrating domain knowledge.

5.2.1. Multimodal integration. Although we limit our review to tabular data, we recognize that using one data modality might capture only a partial view of a patient's health state and potentially lead to suboptimal phenotyping. As highlighted in other reviews (79, 87), the influence of CGMs in medicine would be enhanced by integrating tabular data with complementary modalities, such as free-text clinical notes (86), 3D computed tomography scans (172), or genomic information. Developing generative models for multimodal data generation is a promising area of research, particularly in precision medicine, as digital health twins require simulating multiple sources of patient data.

5.2.2. Incorporating prior knowledge. Our survey focuses on data-driven methods and excludes so-called hybrid generative models that utilize domain-specific knowledge in the form of expert feedback and medical theory (53). While data-driven CGMs have many advantages (e.g., they are easier to train), we recognize that integrating domain expertise and prior biological knowledge could empower CGMs to fully represent the nuances of a patient's physiology. This expertise could be from medical ontologies (173), knowledge graphs (174), or other biological priors in the form of constraints (134, 146, 175).

5.3. Ethical

There is currently a lack of standardized frameworks of how to ethically deploy or use synthetic data. We highlight three such ethical gaps that require our attention: better regulatory oversight, interpretable producer-user contracts, and the prioritization of diversity in generative modeling.

5.3.1. Regulatory oversight. The current relationship between synthetic data and governmental oversight is highly ambiguous, particularly in the United States (48, 78, 176, 177). Outside

of privacy-protecting regulations like HIPAA and a few guidelines for artificial intelligence integration (178, 179), a governmental framework regarding responsible usage of synthetic data is lacking.

Without regulatory oversight, CGMs risk harming patients and compromising the quality of medical research. Similarly to Giuffrè & Shung (78), we emphasize the need for governmental regulations that (*a*) explicitly define synthetic data, (*b*) outline the specific scenarios and extent to which synthetic data can be used to generate evidence, and (*c*) describe how to prevent harm to individuals who contribute their data. Additionally, a regulatory body from diverse backgrounds should be tasked to oversee and enforce these rules.

5.3.2. Interpretable producer–user contract. Currently, computational engineers release a generative model or synthetic dataset, typically along with model evaluation metrics. From these evaluation results, users must determine if the model or data are fit for purpose regarding their specific application. However, as covered in Section 4.3, generative model evaluation is an open problem plagued by a vast number of potentially conflicting metrics and a lack of standardization. Combined with the aforementioned absence of regulatory guidelines, there is a concerning risk of miscommunication and misuse between synthetic data producers and users. To mitigate misuse, we recommend that data producers document acceptable applications of their data (or model) through an interpretable and standardized guideline document. The resulting document for a synthetic dataset might resemble a mixture of Model Cards (180) and Datasheets (181), two existing reporting standards in ML. The scope and content of these guidelines should be established by a diverse group of industry, research, legal, and medical experts.

5.3.3. Prioritizing diversity and fairness. Generative models have historically focused on replicating real data and privacy preservation. However, neglecting diversity and fairness in synthetic data risks harming already disadvantaged populations and misrepresenting rare disease cohorts. For instance, although upsampling via data augmentation is canonically used to correct for representation bias, there is limited research on how augmentation affects diversity and stereotypes.

We thus recommend prioritizing diversity in this field. First, we highlight the need for task-agnostic data diversity metrics, unlike traditional algorithmic fairness metrics that target a specific prediction task. Second, we suggest that all synthetic datasets undergo a standardized and rigorous process evaluating the presence or introduction of harmful biases. Third, we recommend investigating diversity-enhancing CGMs that remove identified harmful biases, as seen by recent efforts like those described in References 71, 182, and 183. Although no panacea, developing generative methods that diversify data could enhance the representation of marginalized groups and promote equity in precision medicine.

SUMMARY POINTS

1. Conditional generative models (CGMs) have the potential to improve the diversity of data representations and advance precision medicine via digital health twins.
2. Our review covers 43 works that propose CGMs for tabular static or time-series data.
3. Although the surveyed works propose state-of-the-art methods to tackle modeling tabular medical data, there remains opportunity for innovation such as the lack of tabular-specific machine learning architectures.

4. Evaluating the quality of synthetic data and generative models is still being actively explored. We highlight five evaluation axes: utility, resemblance, privacy, diversity, and computational cost.

FUTURE ISSUES

1. Conditional generative models (CGMs) for tabular data face technical challenges in modeling complex missingness patterns and interpreting causal mechanisms of generation.
2. Applications of CGMs to medicine would greatly benefit from integrating multimodal data sources and prior biological knowledge.
3. We highlight the need for three developments before CGMs can be ethically adopted: well-defined regulatory oversight, interpretable contracts between producers and users of synthetic data, and prioritizing diversity and fairness assessments.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank Carole Federico for many fruitful conversations around the ethics of synthetic data. We are grateful for the Stanford Human-Centered Artificial Intelligence (HAI) Seed Fund, which helped fund this project.

LITERATURE CITED

1. Johnson KB, Wei W-Q, Weeraratne D, Frisse ME, Misulis K, et al. 2021. Precision medicine, AI, and the future of personalized health care. *Clin. Transl. Sci.* 14(1):86–93
2. Zhang X, Pérez-Stable EJ, Bourne PE, Peprah E, Duru OK, et al. 2017. Big data science: opportunities and challenges to address minority health and health disparities in the 21st century. *Ethn. Dis.* 27(2):95–106
3. Bibbins-Domingo K, Helman A, eds. 2022. Why diverse representation in clinical research matters and the current state of representation within the clinical research ecosystem. In *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups*, pp. 23–46. Washington, DC: Natl. Acad. Press
4. Powell W, Richmond J, Mohottige D, Yen I, Joslyn A, Corbie-Smith G. 2019. Medical mistrust, racism, and delays in preventive health screening among African-American men. *Behav. Med.* 45(2):102–17
5. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative adversarial networks. arXiv:1406.2661 [stat.ML]
6. Ho J, Jain A, Abbeel P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, Vol. 33, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H Lin, pp. 6840–51. Red Hook, NY: Curran Assoc.
7. Yi X, Walia E, Babyn P. 2019. Generative adversarial network in medical imaging: a review. *Med. Image Anal.* 58:101552
8. Kazerouni A, Aghdam EK, Heidari M, Azad R, Fayyaz M, et al. 2023. Diffusion models in medical imaging: a comprehensive survey. *Med. Image Anal.* 88:102846

9. Baowaly MK, Lin C-C, Liu C-L, Chen K-T. 2019. Synthesizing electronic health records using improved generative adversarial networks. *J. Am. Med. Inform. Assoc.* 26(3):228–41
10. van Breugel B, van der Schaar M. 2023. Beyond privacy: navigating the opportunities and challenges of synthetic data. arXiv:2304.03722 [cs.LG]
11. Kingma DP, Welling M. 2013. Auto-encoding variational Bayes. arXiv:1312.6114 [stat.ML]
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 33, ed. U von Luxberg, I Guyon, S Bengio, H Wallach, R Fergus, pp. 6000–10. Red Hook, NY: Curran Assoc.
13. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. 2017. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, Vol. 30, ed. I Guyon, U von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, pp. 5769–79. Red Hook, NY: Curran Assoc.
14. Arjovsky M, Chintala S, Bottou L. 2017. Wasserstein GAN. arXiv:1701.07875 [stat.ML]
15. Mirza M, Osindero S. 2014. Conditional generative adversarial nets arXiv:1411.1784 [cs.LG]
16. Sohl-Dickstein J, Weiss EA, Maheswaranathan N, Ganguli S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. arXiv:1503.03585 [cs.LG]
17. Sherstinsky A. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* 404:132306
18. Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput*. 9(8):1735–80
19. Cho K, van Merriënboer B, Bahdanau D, Bengio Y. 2014. On the properties of neural machine translation: encoder-decoder approaches. arXiv:1409.1259 [cs.CL]
20. Chung J, Kastner K, Dinh L, Goel K, Courville AC, Bengio Y. 2015. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, Vol. 28, ed. C Cortes, N Lawrence, D Lee, M Sugiyama, R Garnett, pp. 2980–88. Red Hook, NY: Curran Assoc.
21. Alammar J. 2018. The illustrated transformer. *Jay Alammar*. **https://jalammar.github.io/illustrated-transformer/**
22. Chokwitthaya C, Zhu Y, Mukhopadhyay S, Jafari A. 2020. Applying the Gaussian mixture model to generate large synthetic data from a small data set. In *Construction Research Congress 2020: Computer Applications*, ed. D Grau, P Tang, M El Asmar, pp. 1251–60. Reston, VA: Am. Soc. Civ. Eng.
23. Oganian A, Domingo-Ferrer J. 2017. Local synthesis for disclosure limitation that satisfies probabilistic *k*-anonymity criterion. *Trans. Data Priv.* 10(1):61–81
24. Emam KE, Mosquera L, Zheng C. 2021. Optimizing the synthesis of clinical trial data using sequential trees. *J. Am. Med. Inform. Assoc.* 28(1):3–13
25. Wang Z, Myles P, Tucker A. 2021. Generating and evaluating cross-sectional synthetic electronic healthcare data: preserving data utility and patient privacy. *Comput. Intell.* 37(2):819–51
26. Young J, Graham P, Penny R. 2009. Using Bayesian networks to create synthetic data. *J. Off. Stat.* 25(4):549–67
27. Kaur D, Sobiesk M, Patil S, Liu J, Bhagat P, et al. 2021. Application of Bayesian networks to generate synthetic health data. *J. Am. Med. Inform. Assoc.* 28(4):801–11
28. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. 2011. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* 20(1):40–49
29. Graham P, Young J, Penny R. 2009. Multiply imputed synthetic data: evaluation of hierarchical Bayesian imputation models. *J. Off. Stat.* 25(2):245–68
30. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16:321–57
31. Xu L, Veeramachaneni K. 2018. Synthesizing tabular data using generative adversarial networks. arXiv:1811.11264 [cs.LG]
32. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. 2017. Generating multi-label discrete patient records using generative adversarial networks. *Proc. Mach. Learn. Res.* 68:286–305
33. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. 2018. Data synthesis based on generative adversarial networks. arXiv:1806.03384 [cs.DB]
34. Arora A, Arora A. 2022. Synthetic patient data in health care: a widening legal loophole. *Lancet* 399(10335):1601–2

35. Beduschi A. 2024. Synthetic data protection: towards a paradigm change in data regulation? *Big Data Soc.* 11(1). **https://doi.org/10.1177/2053951724123**
36. Dwork C. 2006. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ed. M Bugliesi, B Preneel, V Sassone, I Wegener, pp. 1–12. Berlin: Springer
37. Tantipongpipat UT, Waites C, Boob D, Siva AA, Cummings R. 2021. Differentially private synthetic mixed-type data generation for unsupervised learning. *Intell. Decis. Technol.* 15(4):779–807
38. Rosenblatt L, Liu X, Pouyanfar S, de Leon E, Desai A, Allen J. 2020. Differentially private synthetic data: applied evaluations and enhancements. arXiv:2011.05537 [cs.LG]
39. Jordon J, Yoon J, van der Schaar M. 2018. *PATE-GAN: generating synthetic data with differential privacy guarantees*. Paper presented at the Sixth International Conference on Learning Representations, Vancouver, Canada
40. Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X. 2017. PrivBayes: private data release via Bayesian networks. *ACM Trans. Database Syst.* 42(4):25
41. McMahan HB, Moore E, Ramage D, Hampson S, y Arcas BA. 2023. Communication-efficient learning of deep networks from decentralized data. arXiv:1602.05629 [cs.LG]
42. Little C, Elliot M, Allmendinger R. 2023. Federated learning for generating synthetic data: a scoping review. *Int. J. Popul. Data Sci.* 8(1):2158
43. Duan S, Liu C, Han P, Jin X, Zhang X, et al. 2022. HT-Fed-GAN: federated generative model for decentralized tabular data synthesis. *Entropy* 25(1):88
44. Fang ML, Dhami DS, Kersting K. 2022. DP-CTGAN: Differentially private medical data generation using CTGANs. In *Artificial Intelligence in Medicine: 20th International Conference on Artificial Intelligence in Medicine, AIME 2022, Halifax, NS, Canada, June 14–17, 2022, Proceedings*, ed. M Michalowski, SSR Abidi, S Abidi, pp. 178–88. Berlin: Springer
45. Weldon J, Ward T, Brophy E. 2021. Generation of synthetic electronic health records using a federated GAN. arXiv:2109.02543 [cs.LG]
46. Wang Z, Draghi B, Rotalinti Y, Lunn D, Myles P. 2024. High-fidelity synthetic data applications for data augmentation. In *Deep Learning: Recent Findings and Research*, ed. M Domínguez-Morales, J Civit-Masot, L Muñoz-Saavedra, R Damaševičius. London: IntechOpen
47. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K. 2021. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* 11(4):e043497
48. Alloza C, Knox B, Raad H, Aguilà M, Coakley C, et al. 2023. A case for synthetic data in regulatory decision-making in Europe. *Clin. Pharmacol. Ther.* 114(4):795–801
49. OpenSAFELY. 2024. OpenSAFELY: open source platform for NHS electronic health records research. **https://www.opensafely.org/**
50. SMART Health IT. 2024. SMART Health IT: connecting healthcare through open standards. **https://smarthealthit.org/**
51. Laderas T, Vasilevsky N, Pederson B, Haendel M, McWeeney S, Dorr DA. 2017. Teaching data science fundamentals through realistic synthetic clinical cardiovascular data. bioRxiv 232611. **https://doi.org/10.1101/232611**
52. Gonzales A, Guruswamy G, Smith SR. 2023. Synthetic data in health care: a narrative review. *PLOS Digit. Health* 2(1):e0000082
53. Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. 2023. Synthetic data generation: state of the art in health care domain. *Comput. Sci. Rev.* 48:100546
54. James S, Harbron C, Branson J, Sundler M. 2021. Synthetic data use: exploring use cases to optimise data utility. *Discov. Artif. Intell.* 1(1):15
55. Benaim AR, Almog R, Gorelik Y, Hochberg I, Nassar L, et al. 2020. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR Med. Inform.* 8(2):e16492
56. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. 2021. Reproducibility in machine learning for health research: still a ways to go. *Sci. Transl. Med.* 13(586):eabb1655
57. Walonoski J, Klaus S, Granger E, Hall D, Gregorowicz A, et al. 2020. Synthea^{TM} novel coronavirus (COVID-19) model and synthetic data set. *Intell. Based Med.* 1:100007

58. Chen J, Chun D, Patel M, Chiang E, James J. 2019. 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.* 19:44
59. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. 2021. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit. Health* 3(4):e260–65
60. Paik KE, Hicklen R, Kaggwa F, Puyat CV, Nakayama LF, et al. 2023. Digital determinants of health: Health data poverty amplifies existing health disparities—a scoping review. *PLOS Digit. Health* 2(10):e0000313
61. Nazha B, Mishra M, Pentz R, Owonikoko TK. 2019. Enrollment of racial minorities in clinical trials: Old problem assumes new urgency in the age of immunotherapy. *Am. Soc. Clin. Oncol. Educ. Book* 39:3–10
62. US Food Drug Adm. 2020. *Enhancing the diversity of clinical trial populations—eligibility criteria, enrollment practices, and trial designs*. Guidance Doc. FDA-2019-D-1264, US Food Drug Adm., Rockville, MD. **https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enhancing-diversity-clinical-trial-populations-eligibility-criteria-enrollment-practices-and-trial**
63. Kaye JB, Schultz LE, Steiner HE, Kittles RA, Cavallari LH, Karnes JH. 2017. Warfarin pharmacogenomics in diverse populations. *Pharmacotherapy* 37(9):1150–63
64. US Food Drug Adm. 2024. *Diversity action plans to improve enrollment of participants from underrepresented populations in clinical studies*. Draft Guidance Doc. FDA-2021-D-0789, US Food Drug Adm., Rockville, MD. **https://www.fda.gov/regulatory-information/search-fda-guidance-documents/diversity-action-plans-improve-enrollment-participants-underrepresented-populations-clinical-studies**
65. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54(6):115
66. Malygina T, Ericheva E, Drokin I. 2019. Data augmentation with GAN: improving chest X-ray pathologies prediction on class-imbalanced cases. In *Analysis of Images, Social Networks and Texts: 8th International Conference, AIST 2019, Kazan, Russia, July 17–19, 2019, Revised Selected Papers*, ed. WMP van der Aalst, V Batagelj, DI Ignatov, M Khachay, V Kuskova, et al., pp. 321–34. Berlin: Springer
67. Chen Y, Yang XH, Wei Z, Heidari AA, Zheng N, et al. 2022. Generative adversarial networks in medical image augmentation: a review. *Comput. Biol. Med.* 144:105382
68. Ktena I, Wiles O, Albuquerque I, Rebuffi S-A, Tanno R, et al. 2024. Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* 30(4):1166–73
69. Sauber-Cole R, Khoshgoftaar TM. 2022. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *J. Big Data* 9(1):98
70. Choi K, Grover A, Singh T, Shu R, Ermon S. 2020. Fair generative modeling via weak supervision. *Proc. Mach. Learn. Res.* 119:1887–98
71. Um S, Ye JC. 2023. Don't play favorites: minority guidance for diffusion models. arXiv:2301.12334 [cs.LG]
72. Lee J, Kim H, Hong Y, Chung HW. 2021. Self-diagnosing GAN: diagnosing underrepresented samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, Vol. 34, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 1925–38. Red Hook, NY: Curran Assoc.
73. Tan S, Shen Y, Zhou B. 2020. Improving the fairness of deep generative models without retraining. arXiv:2012.04842 [cs.CV]
74. van Breugel B, Kyono T, Berrevoets J, van der Schaar M. 2021. DECAF: Generating fair synthetic data using causally-aware generative networks. In *Advances in Neural Information Processing Systems*, Vol. 34, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 22221–33. Red Hook, NY: Curran Assoc.
75. Pujol D, Gilad A, Machanavajjhala A. 2022. Prefair: privately generating justifiably fair synthetic data. arXiv:2212.10310 [cs.CR]
76. Xu D, Yuan S, Zhang L, Wu X. 2018. Fairgan: fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–75. Piscataway, NJ: IEEE
77. Rajabi A, Garibay OO. 2022. TabFairGAN: fair tabular data generation with generative adversarial networks. *Mach. Learn. Knowl. Extr.* 4(2):488–501

78. Giuffrè M, Shung DL. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Med*. 6(1):186
79. Bordukova M, Makarov N, Rodriguez-Esteban R, Schmich F, Menden MP. 2024. Generative artificial intelligence empowers digital twins in drug discovery and clinical trials. *Expert Opin. Drug Discov.* 19(1):33–42
80. Georges-Filteau J, Cirillo E. 2020. Synthetic observational health data with GANS: from slow adoption to a boom in medical research and ultimately digital twins? arXiv:2005.13510 [cs.LG]
81. Chen J, Yi C, Du H, Niyato D, Kang J, et al. 2024. A revolution of personalized healthcare: enabling human digital twin with mobile AIGC. *IEEE Netw*. **https://doi.org/10.1109/MNET.2024.3366560**
82. Venkatesh KP, Raza MM, Kvedar JC. 2022. Health digital twins as tools for precision medicine: considerations for computation, implementation, and regulation. *NPJ Digit. Med.* 5(1):150
83. Myles P, Ordish J, Tucker A. 2023. The potential synergies between synthetic data and *in silico* trials in relation to generating representative virtual population cohorts. *Prog. Biomed. Eng.* 5(1):013001
84. Sanson-Fisher RW, Bonevski B, Green LW, D'Este C. 2007. Limitations of the randomized controlled trial in evaluating population-based health interventions. *Am. J. Prev. Med.* 33(2):155–61
85. Ghosheh GO, Li J, Zhu T. 2024. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Comput. Surv.* 56(6):147
86. Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, et al. 2024. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit. Health* 6(4):e281–90
87. Ghebrehiwet I, Zaki N, Damseh R, Mohamad MS. 2024. Revolutionizing personalized medicine with generative AI: a systematic review. *Artif. Intell. Rev.* 57:128
88. Yoon J, Jordon J, van der Schaar M. 2018. *GANITE: estimation of individualized treatment effects using generative adversarial nets*. Paper presented at the Sixth International Conference on Learning Representations, Vancouver, Canada
89. Komanduri A, Wu X, Wu Y, Chen F. 2024. From identifiable causal representations to controllable counterfactual generation: a survey on causal generative modeling. arXiv:2310.11011v2 [cs.LG]
90. Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, et al. 2022. Synthetic data – what, why and how? arXiv:2205.03257 [cs.LG]
91. Figueira A, Vaz B. 2022. Survey on synthetic data generation, evaluation methods and GANs. *Mathematics* 10(15):2733
92. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. 2020. Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* 20:108
93. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. 2022. Synthetic data generation for tabular health records: a systematic review. *Neurocomputing* 493:28–45
94. Perkonoja K, Auranen K, Virta J. 2023. Methods for generating and evaluating synthetic longitudinal patient data: a systematic review. arXiv:2309.12380 [stat.ME]
95. Xing X, Wu H, Wang L, Stenson I, Yong M, et al. 2023. Non-imaging medical data synthesis for trustworthy AI: a comprehensive survey. *ACM Comput. Surv.* 56(7):165
96. Liu T, Fan J, Li G, Tang N, Du X. 2024. Tabular data synthesis with generative adversarial networks: design space and optimizations. *VLDB J.* 33(2):255–80
97. Ma F, Wang Y, Gao J, Xiao H, Zhou J. 2020. Rare disease prediction by generating quality-assured electronic health records. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, ed. C Demeniconi, N Chawla, pp. 514–22. Philadelphia, PA: Soc. Ind. Appl. Math.
98. Yang Z, Guo P, Zanna K, Sano A. 2024. Balanced mixed-type tabular data synthesis with diffusion models. arXiv:2404.08254 [cs.LG]
99. Lu C, Reddy CK, Wang P, Nie D, Ning Y. 2023. Multi-label clinical time-series generation via conditional GAN. *IEEE Trans. Knowl. Data Eng.* 36(4):1728–40
100. Engelmann J, Lessmann S. 2021. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. *Expert Syst. Appl.* 174:114582
101. Liu T, Fan J, Tang N, Li G, Du X. 2024. Controllable tabular data synthesis using diffusion models. In *Proceedings of the ACM on Management of Data*, Vol. 2, ed. D Agrawal, Art. 28. New York: Assoc. Comput. Mach.

102. Bing S, Dittadi A, Bauer S, Schwab P. 2022. Conditional generation of medical time series for extrapolation to underrepresented populations. *PLOS Digit. Health* 1(7):e0000074
103. Ramachandranpillai R, Sikder MF, Bergström D, Heintz F. 2024. Bt-GAN: generating fair synthetic healthdata via bias-transforming generative adversarial networks. *J. Artif. Intell. Res.* 79:1313–41
104. Liu Z, Li X, Philip SY. 2023. A counterfactual fair model for longitudinal electronic health records via Deconfounder. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1175–80. Piscataway, NJ: IEEE
105. Liu T, Qian Z, Berrevoets J, van der Schaar M. 2022. *GOGGLE: generative modelling for tabular data by learning relational structure*. Paper presented at the Eleventh International Conference on Learning Representations, Kigali, Rwanda
106. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. 2019. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, Vol. 32, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett, pp. 7335–45. Red Hook, NY: Curran Assoc.
107. Abroshan M, Elliott A, Khalili MM. 2024. Imposing fairness constraints in synthetic data generation. *Proc. Mach. Learn. Res.* 238:2269–77
108. Zhao Z, Kunar A, Birke R, Van der Scheer H, LY Chen. 2024. CTAB-GAN+: enhancing tabular data synthesis. *Front. Big Data* 6:1296508
109. Tan Y, Zhu H, Wu J, Chai H. 2024. DPTVAE: data-driven prior-based tabular variational autoencoder for credit data synthesizing. *Expert Syst. Appl.* 241:122071
110. Theodorou B, Xiao C, Sun J. 2023. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nat. Commun.* 14(1):5305
111. Ullah S, Xu Z, Wang H, Menzel S, Sendhoff B, Bäck T. 2020. Exploring clinical time series forecasting with meta-features in variational recurrent models. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. Piscataway, NJ: IEEE
112. Bhave S, Perotte A. 2021. Point processes for competing observations with recurrent networks (POPCORN): a generative model of EHR data. *Proc. Mach. Learn. Res.* 149:770–89
113. Rasul K, Seward C, Schuster I, Vollgraf R. 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. *Proc. Mach. Learn. Res.* 139:8857–68
114. Biswal S, Ghosh S, Duke J, Malin B, Stewart W, et al. 2021. EVA: generating longitudinal electronic health records using conditional variational autoencoders. *Proc. Mach. Learn. Res.* 149:260–82
115. Teng X, Pei S, Lin Y-R. 2020. StoCast: stochastic disease forecasting with progression uncertainty. *IEEE J. Biomed. Health Inform.* 25(3):850–61
116. Das T, Wang Z, Sun J. 2023. TWIN: personalized clinical trial digital twin generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 402–13. New York: Assoc. Comput. Mach.
117. Walsh JR, Smith AM, Pouliot Y, Li-Bland D, Loukianov A, Fisher CK. 2020. Generating digital twins with multiple sclerosis using probabilistic neural networks. arXiv:2002.02779 [stat.ML]
118. Li J, Cairns BJ, Li J, Zhu T. 2023. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digit. Med.* 6(1):98
119. Sun Z, Sun Z, Dong W, Shi J, Huang Z. 2021. Towards predictive analysis on disease progression: a variational Hawkes process model. *IEEE J. Biomed. Health Inform.* 25(11):4195–206
120. Dedhia B, Balasubramanian R, Jha NK. 2023. SCouT: synthetic counterfactuals via spatiotemporal transformers for actionable healthcare. *ACM Trans. Comput. Healthc.* 4(4):23
121. Zhang Z, Yan C, Lasko TA, Sun J, Malin BA. 2021. SynTEG: a framework for temporal structured electronic health data simulation. *J. Am. Med. Inform. Assoc.* 28(3):596–604
122. Shankar V, Yousefi E, Manashty A, Blair D, Teegapuram D. 2023. Clinical-GAN: trajectory forecasting of clinical events using transformer and generative adversarial networks. *Artif. Intell. Med.* 138:102507
123. Pang C, Jiang X, Pavinkurve NP, Kalluri KS, Minto EL, et al. 2024. CEHR-GPT: generating electronic health records with chronological patient timelines. arXiv:2402.04400 [cs.LG]
124. Lee JM, Hauskrecht M. 2021. Modeling multivariate clinical event time-series with recurrent temporal mechanisms. *Artif. Intell. Med.* 112:102021
125. Li L, Yan J, Zhang Y, Zhang J, Bao J, et al. 2022. Learning generative RNN-ODE for collaborative time-series and event sequence forecasting. *IEEE Trans. Knowl. Data Eng.* 35(7):7118–37

126. Wang Z, Sun J. 2022. PromptEHR: conditional electronic healthcare records generation with prompt learning. arXiv:2211.01761 [cs.CL]
127. Mosquera L, El Emam K, Ding L, Sharma V, Zhang XH, et al. 2023. A method for generating synthetic longitudinal health data. *BMC Med. Res. Methodol.* 23(1):67
128. Guyomard V, Fessant F, Guyet T, Bouadi T, Termier A. 2022. VCNet: a self-explaining model for realistic counterfactual generation. In *Machine Learning and Knowledge Discovery in Databases*, ed. MR Amini, S Canu, A Fischer, T Guns, P Kralj Novak, G Tsoumakas, pp. 437–53. Cham, Switz.: Springer
129. Samoilescu R-F, Van Looveren A, Klaise J. 2021. Model-agnostic and scalable counterfactual explanations via reinforcement learning. arXiv:2106.02597 [cs.LG]
130. Duong TD, Li Q, Xu G. 2023. CeFlow: a robust and efficient counterfactual explanation framework for tabular data using normalizing flows. In *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Osaka, Japan, May 25–28, 2023, Proceedings, Part II*, ed. H Kashima, T Ide, W-C Peng, pp. 133–44. Cham, Switz.: Springer
131. Nemirovsky D, Thiebaut N, Xu Y, Gupta A. 2022. CounteRGAN: generating counterfactuals for real-time recourse and interpretability using residual GANs. *Proc. Mach. Learn. Res.* 180:1488–97
132. Rashidian S, Wang F, Moffitt R, Garcia V, Dutt A, et al. 2020. SMOOTH-GAN: towards sharp and smooth synthetic EHR data generation. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings*, ed. M Michalowski, R Moskovitch, pp. 37–48. Cham, Switz.: Springer
133. Sun S, Wang F, Rashidian S, Kurc T, Abell-Hart K, et al. 2021. Generating longitudinal synthetic EHR data with recurrent autoencoders and generative adversarial networks. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB Workshops, Poly 2021 and DMAH 2021, Virtual Event, August 20, 2021, Revised Selected Papers 7*, pp. 153–65. Cham, Switz.: Springer
134. Han P, Xu W, Lin W, Cao J, Liu C, et al. 2023. C3-TGAN-controllable tabular data synthesis with explicit correlations and property constraints. Authorea Preprint. **https://www.techrxiv.org/doi/full/10.36227/techrxiv.24249643.v1**
135. Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. 2023. TabDDPM: modelling tabular data with diffusion models. *Proc. Mach. Learn. Res.* 202:17564–79
136. Solatorio AV, Dupriez O. 2023. REaLTabFormer: generating realistic relational and tabular data using transformers. arXiv:2302.02041 [cs.LG]
137. Borisov V, Seßler K, Leemann T, Pawelczyk M, Kasneci G. 2022. Language models are realistic tabular data generators. arXiv:2210.06280 [cs.LG]
138. Gulati M, Roysdon P. 2024. TabMT: generating tabular data with masked transformers. In *Advances in Neural Information Processing Systems*, Vol. 36, ed. A Oh, T Naumann, A Globerson, K Saenko, M Hardt, S Levine, pp. 46245–54. Red Hook, NY: Curran Assoc.
139. Baak M, Brugman S, Rojas IF, Dalmeida L, Urlus RE, Oger JB. 2022. Synthsonic: fast, probabilistic modeling and synthesis of tabular data. *Proc. Mach. Learn. Res.* 151:4747–63
140. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–26
141. Wang Y, Blei DM. 2019. The blessings of multiple causes. *J. Am. Stat. Assoc.* 114(528):1574–96
142. Kusner MJ, Loftus J, Russell C, Silva R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, Vol. 30, ed. I Guyon, U von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, pp. 4069–79. Red Hook, NY: Curran Assoc.
143. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, et al. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461 [cs.CL]
144. Shchur O, Türkmen AC, Januschowski T, Günnemann S. 2021. Neural temporal point processes: a review. arXiv:2104.03528 [cs.LG]
145. Pearl J. 2019. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* 62(3):54–60
146. Stoian MC, Dyrmishi S, Cordy M, Lukasiewicz T, Giunchiglia E. 2024. How realistic is your synthetic data? Constraining deep generative models for tabular data. arXiv:2402.04823 [cs.LG]

147. Hoogeboom E, Nielsen D, Jaini P, Forré P, Welling M. 2021. Argmax flows and multinomial diffusion: learning categorical distributions. In *Advances in Neural Information Processing Systems*, Vol. 34, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 12454–65. Red Hook, NY: Curran Assoc.
148. Groenwold RHH. 2020. Informative missingness in electronic health record systems: the curse of knowing. *Diagn. Progn. Res.* 4(1):8
149. Nijman S, Leeuwenberg A, Beekers I, Verkouter I, Jacobs J, et al. 2022. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J. Clin. Epidemiol.* 142:218–29
150. Nong P, Williamson A, Anthony D, Platt J, Kardia S. 2022. Discrimination, trust, and withholding information from providers: implications for missing data and inequity. *SSM Popul. Health* 18:101092
151. Jang E, Gu S, Poole B. 2016. Categorical reparameterization with Gumbel-Softmax. arXiv:1611.01144 [stat.ML]
152. Ceritli T, Ghosheh GO, Chauhan VK, Zhu T, Creagh AP, Clifton DA. 2023. Synthesizing mixed-type electronic health records using diffusion models. arXiv:2302.14679 [cs.LG]
153. Müller-Franzes G, Niehues JM, Khader F, Arasteh ST, Haarburger C, et al. 2023. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Sci. Rep.* 13(1):12098
154. Dhariwal P, Nichol A. 2021. Diffusion models beat GANS on image synthesis. In *Advances in Neural Information Processing Systems*, Vol. 34, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 8780–94. Red Hook, NY: Curran Assoc.
155. Dieleman S, Sartran L, Roshannai A, Savinov N, Ganin Y, et al. 2022. Continuous diffusion for categorical data. arXiv:2211.15089 [cs.CL]
156. Radford A, Wu J, Child R, Luan D, Amodei D, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*. **https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf**
157. Lee J, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL]
158. Alaa A, Van Breugel B, Saveliev ES, van der Schaar M. 2022. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. *Proc. Mach. Learn. Res.* 162:290–306
159. Hernadez M, Epelde G, Alberdi A, Cilla R, Rankin D. 2023. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods Inf. Med.* 62(Suppl. 1):e19–38
160. Vallevik VB, Babic A, Marshall SE, Elvatun S, Brøgger H, et al. 2024. Can I trust my fake data – a comprehensive quality assessment framework for synthetic tabular data in healthcare. arXiv:2401.13716 [cs.LG]
161. Du Y, Li N. 2024. Systematic assessment of tabular data synthesis algorithms. arXiv:2402.06806 [cs.CR]
162. Pereira M, Kshirsagar M, Mukherjee S, Dodhia R, Lavista Ferres J, de Sousa R. 2024. Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *PLOS ONE* 19(2):e0297271
163. Yang SC-H, Eaves B, Schmidt M, Swanson K, Shafto P. 2024. Structured evaluation of synthetic tabular data. arXiv:2403.10424 [cs.LG]
164. Naeem MF, Oh SJ, Uh Y, Choi Y, Yoo J. 2020. Reliable fidelity and diversity metrics for generative models. *Proc. Mach. Learn. Res.* 119:7176–85
165. Qian Z, Davis R, van der Schaar M. 2024. Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. In *Advances in Neural Information Processing Systems*, Vol. 36, ed. A Oh, T Naumann, A Globerson, K Saenko, M Hardt, S Levine, pp. 3173–88. Red Hook, NY: Curran Assoc.
166. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. 2021. The problem of fairness in synthetic healthcare data. *Entropy* 23(9):1165
167. Chen J, Xu Y, Wang P, Yang Y. 2023. Deep generative imputation model for missing not at random data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 316–25. New York: Assoc. Comput. Mach.

168. Pereira RC, Abreu PH, Rodrigues PP. 2022. Partial multiple imputation with variational autoencoders: tackling not at randomness in healthcare data. *IEEE J. Biomed. Health Inform.* 26(8):4218–27
169. Ma C, Zhang C. 2021. Identifiable generative models for missing not at random data imputation. In *Advances in Neural Information Processing Systems*, Vol. 34, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 27645–58. Red Hook, NY: Curran Assoc.
170. Wen B, Cao Y, Yang F, Subbalakshmi K, Chandramouli R. 2022. *Causal-TGAN: modeling tabular data using causally-aware GAN*. Poster presented at the ICLR Workshop on Deep Generative Models for Highly Structured Data, Apr. 29. **https://iclr.cc/virtual/2022/7866**
171. Rahman MM, Jordan M, Kocaoglu M. 2024. Conditional generative models are sufficient to sample from any causal effect estimand. arXiv:2402.07419 [cs.LG]
172. Ziegler JD, Subramaniam S, Azzarito M, Doyle O, Krusche P, Coroller T. 2022. *Multi-modal conditional GAN: data synthesis in the medical domain*. Poster presented at the NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, Dec. 2. **https://neurips.cc/virtual/2022/58678**
173. Niu S, Ma J, Bai L, Wang Z, Guo L, Yang X. 2024. EHR-KnowGen: knowledge-enhanced multimodal learning for disease diagnosis generation. *Inform. Fusion* 102:102069
174. Li R, Yin C, Yang S, Qian B, Zhang P. 2020. Marrying medical domain knowledge with deep learning on electronic health records: a deep visual analytics approach. *J. Med. Internet Res.* 22(9):e20645
175. Theodorou B, Jain S, Xiao C, Sun J. 2024. ConSequence: synthesizing logically constrained sequences for electronic health record generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, ed. M Wooldridge, J Dy, S Natarajan, pp. 15355–63. Washington, DC: AAAI
176. Off. Shared Solut. Perform. Improv. (OSSPI), Chief Data Off. Counc. (CDO). 2024. Request for information—synthetic data generation. *Federal Register* 89(4):783–86. **https://www.govinfo.gov/content/pkg/FR-2024-01-05/pdf/2024-00036.pdf**
177. Beduschi A. 2024. Synthetic data protection: towards a paradigm change in data regulation? *Big Data Soc.* 11(1). **https://doi.org/10.1177/20539517241231277**
178. Food Drug Adm US. 2024. Artificial intelligence and machine learning (AI/ML) software as a medical device. *US Food and Drug Administration*. **https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device**
179. White House Off. Sci. Technol. Policy. 2024. Blueprint for an AI Bill of Rights. **https://www.whitehouse.gov/ostp/ai-bill-of-rights/**
180. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, et al. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–29. New York: Assoc. Comput. Mach.
181. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, et al. 2021. Datasheets for datasets. *Commun. ACM* 64(12):86–92
182. Madras D, Creager E, Pitassi T, Zemel R. 2019. Fairness through causal awareness: learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 349–58. New York: Assoc. Comput. Mach.
183. Parihar R, Bhat A, Basu A, Mallick S, Kundu JN, Babu RV. 2024. Balancing act: distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6668–78. Piscataway, NJ: IEEE