Synthetic Observational Health Data with GANs: a Cambrian radiation in medical research and digital twins?

Jeremy Georges-Filteau[1,2] and Elisa Cirillo[2]

[1]Radboud University Nijmegen
[2]The Hyve

October 12, 2020

Abstract

After being collected for patient care, Observational Health Data (OHD) can further benefit patient well-being by sustaining the development of health informatics and medical research. Vast potential is largely unexploited due to the fiercely private nature of patient-related data and regulation about its distribution. Generative Adversarial Network (GAN) have recently emerged as a groundbreaking approach to efficiently learn generative models that produce realistic Synthetic Data (SD). They have revolutionized practices in multiple domains such as cars, fraud detection, simulations in the industrial sector and marketing known as digital twins, and medical imaging. The digital twin concept could readily be applied to modelling and quantifying disease progression. In addition, GANs posses a multitude of capabilities that are directly applicable to common problems in the healthcare: augmenting small dataset, correcting class imbalance, domain translation for rare diseases, let alone preserving privacy. Unlocking open access to privacy-preserving OHD could be transformative for scientific research. In the midst of the COVID-19 pandemic, the healthcare system is facing unprecedented challenges, many of which of are data related and cloud be alleviated by the capabilities of GANs. In light of these facts, publications concerning the development of GAN applied to OHD seemed to be severely lacking. To uncover the reasons for the slow adoption of GANs for OHD, we conducted a broad review of the published literature on the subject. Our findings show that the properties of OHD and evaluating the SD were initially challenging for the existing GAN algorithms and metrics (unlike medical imaging, for which state-of-the-art model were directly transferable). Nonetheless, since 2017 solutions to these problems are being published at an increasing rate.

1 Introduction

1.1 Background

Medical professionals collect Observational Health Data (OHD) in Electronic Health Records (EHRs) at various points of care in a patient's trajectory, to support and enable their work (Cowie et al., 2016). The patient profiles found in EHRs are diverse and longitudinal, composed of demographic variables, recordings of diagnoses, conditions, procedures, prescriptions, measurements and lab test results, administrative information, and increasingly omics (OHDSI, 2020).

Having served its primary purpose, this wealth of detailed information can further benefit patient well-being by sustaining medical research and development. That is to say, improving the development life-cycle of Health Informatics (HI), the predictive accuracy of Machine Learning (ML) algorithms, or enabling discoveries in research on clinical decisions, triage decisions, inter-institution collaboration and HI automation (Rudin et al., 2020). Big health data is the underpinning of two prime objectives of precision medicine: individualization of patient interventions and inferring the workings of biological systems from high-level analysis (Capobianco, 2020). However, the private nature of patient-related

data, and the growing widespread concern over its disclosure, hampers dramatically the potential for secondary usage of OHD for legitimate purposes.

Anonymization techniques are used to hinder the misuse of sensitive data. This implies a costly and data-specific cleansing process, and the unavoidable trade-off of enhancing privacy to the detriment of data utility. These techniques are fallible and do not prevent reidentification. In fact, it has been demonstrated that no polynomial time Differential privacy (DP) algorithms can produce Synthetic Data (SD) preserving all relations of the real data, even for simple relations such as 2-way marginals (Ullman and Vadhan, 2011). To address these drawbacks, alternative modes for sharing sensitive data is an active research area, including privacy-preserving analytic and distributed learning. Although promising, these approaches come with limitations and their feasibility has yet to be demonstrated. Regardless, distributed models are vulnerable to a variety of attacks, for which no single protection measure is sufficient as research on defense is far behind attack (Enthoven and Al-Ars, 2020; Gao et al., 2020).The process of DP may also

These conditions restrict access to OHD to professionals with academic credentials and financial resources. The use of OHD by all other health data-related occupations is blocked, along with the downstream benefits. For example, software developers rarely have access to the data at the core of the HI solutions they are developing.

1.2 Synthetic data

An alternative to traditional privacy-preserving methods is to produce full SD with methods categorized as either theory-driven (theoretical, mechanistic or iconic) and data-driven (empirical or interpolatory) modelling (Kim et al., 2017; Hand, 2019). Theory-driven modelling involves a complex knowledge-based attempt to define a simulation process representing the causal relationships of a system, it's mechanism. The Synthea (Walonoski et al., 2017) synthetic patient generator is one such model, in which state transition models[1] produce patient trajectories. The model parameters are taken from aggregate population-level statistics of disease progression and medical knowledge. Such a knowledge-based model depends on prior knowledge of the system, and most importantly how much we can intellect about it (Kim et al., 2017). On one hand theory-based modelling aims at understanding and offers interpretability, on the other when modelling complex systems, simplifications and assumptions are inevitable, leading to inaccuracies (Hand, 2019). In fact, relying on population-level statistics does not produce models capable of reproducing heterogeneous health outcomes (Chen et al., 2019a).

Data-driven modelling techniques infer a representation of the data from a sample distribution, in an attempt to summarize or describe it (Hand, 2019). There exist numerous statistical modelling approaches to produce SD, but the techniques are based on intrinsic assumptions about the data. The representational power is bound to correlations that are intelligible to the modeler, being prone to obscure inaccuracies. SD generated by these models tends to hit a ceiling of utility (Rankin et al., 2020). In the ML field, generative models learn an approximation of the multi-modal distribution, from which synthetic samples can be drawn (Goodfellow, 2016). Generative Adversarial Network (GAN) (Goodfellow et al., 2014) have recently emerged as a groundbreaking approach to efficiently learn generative models that produce realistic SD using Neural Network (NN). GAN algorithms have rapidly found a wide range of applications, such as data augmentation in medical imaging (Yi et al., 2019a; Wang et al., 2020a; Zhou et al., 2020).

The potential impacts of GAN to healthcare and science are considerable, some of which have been realized in fields such as medical imaging. However, the application of GAN to OHD seems to have been lagging (Xiao et al., 2018a). Certain characteristics of OHD could serve to explain the relatively slow progress. Primarily, algorithms developed for images and text in other fields were easily re-purposed for medical equivalents of the data types. However, OHD presents a unique complexity in terms of multi-modality, heterogeneity, and fragmentation (Xiao et al., 2018a). In addition to this, evaluating the

[1] Probabilistic model composed of predefined states, transitions, and conditional logic.

realism of synthetic OHD is intuitively complex, a problem that still burdens GAN in general. Nonetheless, in 2017 the first few attempts at GANs for OHD were published (Esteban et al., 2017; Che et al., 2017; Choi et al., 2017a; Yahi et al., 2017). We aimed to investigate if the field continued to expand following these first few examples, and if so to gain an comprehensive understanding of methods and approaches to the problem.

2 Methods

Table 1: Search query terms

Health data			Generative adversarial models	
	Terms			Terms
OR	clinical health EHR electronic health record patient	AND	OR	generative adversarial GAN adversarial training synthetic

Publications concerning GANs for Observation Health Data (OHD-GAN) were identified through with Google Scholar (Google), Web of Science (Clarivate) and Prophy (Prophy). The search input was formed from the terms and operators found in Table 2. We included studies reporting the development, application, performance evaluation and privacy evaluation of GAN algorithms to produce OHD. Broadly, we define the scope of OHD as categorical, real-valued, ordinal or binary event data recorded for patient care. A more detailed summary of the included and excluded data types can be found in Table 3. The data types are already the subject of one or more review, or would merit a review of their own (Yi et al., 2019b; Nakata, 2019; Anwar et al., 2018; Wang et al., 2020a; Zhou et al., 2020). In each of the included publications, we considered the aspects listed in Table 1.

Table 2: Aspects analysed in each of the publications included in the review

A) Types of healthcare data	D) Evaluation metrics
B) GAN algorithm, learning procedures, losses	E) Privacy considerations
C) Intended use of the SD	F) Interpreatability of the model

Table 3: Types of OHD data included or excluded from the review.

Type	Examples	
Included	Observations	Demographic information, medical classification, family history
	Timestamped observations	Diagnosis, treatment and procedure codes, prescription and dosage, laboratory test results, physiologic measurements and intake events
	Encounters	Visit dates, care provider, care site
	Derived	Aggregated counts, calculated indicators.
Excluded	Omics	Genome, transcriptome, proteome, immunome, metabolome, microbiome
	Imaging	X-rays, computed tomography (CT), magnetic resonance imaging (MRI)
	Signal	Electrocardiogram (ECG), electroencephalogram (EEG)
	Unstructured	Narrative reports, textual

3 Results

3.1 Summary

We have found a total of 43 publications describing the development or adaption of OHD-GAN, presented in Table 4. The type of data addressed in each of these publications can be generalized into

one of two categories: time-dependent observations, such as time-series, or static representation in the form of feature vectors such as tabular rows.

Most efforts propose adaptations of current algorithms to the characteristics and complexities of OHD. These include multi-modality of marginal distributions or non-Gaussian continuous features, heterogeneity, a combination of discrete and continuous features, longitudinal irregularity, complex conditional distributions, missingness or sparsity, class imbalance of categorical features and noise.

While these properties may make training a useful model difficult, the variety of applications that are highly relevant and needed in the healthcare domain provide sufficient incentive. The most cited motives are, as one would expect, to cope with the often limited number of samples in medical datasets and to overcome the highly restricted access to OHD. The potential of releasing privacy-preserving SD freely is a common subject. Publications considering privacy evaluate the effect on utility of applying DP to their algorithm, propose alternatives privacy concepts and metrics, or exclusively concentrate on the subject of privacy.

3.2 Motives for developing OHD-GAN

Some claim that the ability to generate synthetic is becoming an essential skill in data science (Sarkar, 2018), but what purpose can it serve in the medical domain? The authors mention a wide range of potential applications. We briefly describe the four prevailing themes in the following sections: data augmentation (Sec.3.2.1), privacy and accessibility (Sec.3.2.2), precision medicine (Sec.3.2.3) and modelling simulations (Sec.3.2.4).

3.2.1 Data augmentation

Data augmentation is mentioned in nearly all publications. Although counter-intuitive, it is well known that GAN can generate SD that conveys more information about the real data distribution. Effectively, the continuous space distribution of the generator produces a more comprehensive set of data points, valid, but not present in the discrete real data points. A combination of real and synthetic training data habitually leads to increased predictor performance (Wang et al., 2019; Che et al., 2017; Yoon et al., 2018a,b; Yang et al., 2019a; Chen et al., 2019a; Cui et al., 2019; Che et al., 2017). A more intelligible way to seize the concept from the point of view of image classification, in which it is known as invariances, perturbations such as rotation, shift, sheer and scale (Antoniou et al., 2017).

Similarly, domain translation and Semi-supervised learning (SSL) training approaches with GANs could support predictive tasks that lack data with accurate labels, lack paired samples or suffer class imbalance (Che et al., 2017; McDermott et al., 2018; Yoon et al., 2018a). Another example is correcting discrepancies between datasets collected in different locations or under different conditions inducing bias (Yoon et al., 2018c). GANs are also well adapted for data imputation, were entries are Missing at Random (MaR) (Yoon et al., 2018b).

3.2.2 Enhancing privacy and increasing data accessibility

SD is seen as the key to unlocking the unexploited value of OHD hindering machine learning, and more generally scientific progress (Beaulieu-Jones et al., 2019; Baowaly et al., 2019; Baowaly et al., 2018; Che et al., 2017; Esteban et al., 2017; Fisher et al., 2019; Severo et al., 2019). Preserving privacy can broadly be described as reducing the risk of reidentification attack to an acceptable level. This level of risk is quantified when releasing data anonymized with DP.

Due to its artificial nature, SD is put forward as a means to forgo the tight restrictions on data sharing, while potentially providing greater privacy guarantees (Beaulieu-Jones et al., 2019; Baowaly et al., 2019; Baowaly et al., 2018; Esteban et al., 2017; Fisher et al., 2019; Walsh et al., 2020; Chin-Cheong et al., 2019). Enabling access to greater variety, quality and quantity of OHD could have positive effects in a wide range of fields, such as software development, education, and training of medical professionals. The fact remains that GANs do not eliminate the risk of reidentification. Considering none of the synthetic data points represent real people, the significance of such an occurrence is unclear. Nonetheless, both

methods can be combined, and GAN training according to DP shows evidence of reducing the loss of utility in comparison to DP alone. Overall,

3.2.3 Enabling precision medicine

The application to precision medicine generally involve predicting outcomes conditioned on a patient's current state and history. Simulated trajectories could help inform clinical decision making by quantifying disease progression and outcomes and have a transformative effect on healthcare (Walsh et al., 2020; Fisher et al., 2019). Ensembles of stochastic simulations of individual patient profiles such as those produced by Conditional Restricted Boltzmann Machine (CRMB) could help quantify risk at an unprecedented level of granularity (Fisher et al., 2019).

Predicting patient-specific responses to drugs is still a new field of research, a problem known as Individualized Treatment effects (ITE). The task of estimating ITE is persistently hampered by the lack of paired counterfactual samples (Yoon et al., 2018a; Chu et al., 2019). To solve similar problems n medical imaging, various GAN algorithms were developed for domain translation, mapping a sample from its to original class to the paired equivalent. This includes bidirectional transformations, allowing GAN to learn mappings from very few, or a lack of paired samples (Wolterink et al., 2017; Zhu et al., 2017a; McDermott et al., 2018).

3.2.4 From patient and disease models to digital twins

A well trained model approximates the process that generated the real data points. In other words, the relations learned by the model, its parameters, contains meaningful information if we can learn to harness it. Data-driven algorithms evolve as our understanding of their behavior improves. New concepts are incorporated in the algorithms leading to further understanding, iterativly blurring the line with theory-driven approaches (Hand, 2019). Interpretability is a growing field of research concerned with understanding how the learned parameters of a model relate. In other words analysing the representation the algorithm has converged to and deriving meaning from seemingly obscure logic.Incorporating new understanding in the architecture of algorithms shift the view from a data-driven to a theory-driven perspective (Hand, 2019). As we purposefully build structure in our algorithms from new understanding we may get the chance to explore meaningful representations that would otherwise be beyond our reasoning.

Approaching these ideas from above, the concept of "digital twins" represents in a way the ultimate realization of Personalized Medicine. A common practice in industrial sectors is high-fidelity virtual representations of physical assets. Long-term simulations, that provide an overview and comprehensive understanding of the workings, behavior and life-cycle of their real counterparts. The state of the models is continuously updated from theoretical data, real data and streaming Internet of Things (IoT) indicators.

Intently conditioned input data allows the exploration of specific events or conditions. In a position paper on the subject, Angulo et al. draw the parallels of this technique with the current needs in healthcare and the emergence of the necessary technologies for actionable models of patients. (Angulo Bahon et al., 2019; Angulo et al., 2020). The authors bring up the rapid adoption of wearables that are continuously monitoring people's physiological state. Wearables are one of many mobile digitally connected devices that collect patient data over a broad range of physiological characteristic and behavioral patterns (Coravos et al., 2019). This emerging trend known as digital bio-markers has already led to studies demonstrating predictive models with the potential for improved patient care (Snyder et al., 2018). Through continuous lifelong learning, integrating multiple modes of personal data, generative patient models could inform diagnostics of medical professionals and also enable testing treatment options. In their proposal, GAN are an essential component of the ecosystem to ensure patient privacy and to provide bootstrap data. Fisher et al. already employ the term "digital twin" to describe their process, noting that they present no privacy risk and enable simulating patient cohorts of any size and characteristics (Walsh et al., 2020).

Table 4: Summary of the publication included in the review.

Publication	Algorithm(s)	Focus and topics	Other algorithms and topics	Data type
2017				
Choi et al.	medGAN (medGAN)	Incompatibility of back-propagation with discrete features.	Autoencoder (AE), Mini-batch Averaging (MB-Avg), batch-normalization (BN), shortcut connections (SC), Attribute Disclosure (AD), Presence Disclosure (PD)	Discrete features, aggregated counts and occurrences of medical codes.
Yahi et al.	medGAN adaptation	Drug Laboratory Effects (DLE) on continuous time-series, multi-modality.	t-Distributed Stochastic Neighbor Embedding (t-SNE).	Paired pre/post treatment exposure time-series
Esteban et al.	Recurrent GAN (RGAN), Recurrent Convolutional GAN (RC-GAN)	Adversarial training of (conditional) Recurrent NNs (RNNs) on time-series, evaluation, privacy.	Long Short-term Memory (LSTM), Conditional GAN (CGAN), Differential private stochastic gradient descent (DP-SGD)	Regularly observed real-valued time-series
Xiao et al.	WGAN for Temporal Point-processes (WGANTPP)	Temporal Point Processes.	LSTM, Wassertein GAN (WGAN), Poisson process	Sporadic occurrences, hospital visits.
Che et al.	Electronic Health Record GAN (ehrGAN), Semi-supervised Learning with a learned ehrGAN (SSL-GAN)	Semi-supervised augmentation, transitional distribution.	1D-CNN, Word2vec, Variational contrastive divergence (VCD)	Discrete time-series, sequences of medical codes.
Dash et al.	(HealthGAN)	Sleep patterns, stratification by covariates.	-	Binary sleep data obtained by transformation of multiple visits.
2018				
Camino et al.	Multi-categorical ARAE (MC-ARAE), Multi-categorical medGAN (MC-medGAN), Multi-categorical Gumbel-softmax GAN (MC-GumbelGAN), Multi-categorical WGAN with Gradient Penalty (MC-WGAN-GP)	Improving training process.	medGAN, WGAN with Gradient Penalty (WGAN-GP), Gumbel-Softmax GAN (Gumbel-GAN), Adversarially regularized autoencoder (ARAE)	Data composed of multiple categorical variables, represented as one-hot encoded vectors.
McDermott et al.	Cycle Wasserstein Regression GAN (CWR-GAN)	Cycle-consistent semi-supervised regression learning, unpaired data, class imbalance.	WGAN Cycle-consistent GAN (Cycle-GAN) ITE	ICU time-series with lack of paired samples, SD.
Yoon et al.	Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE)	ITE, unobserved counterfactual, multi-label classification, uncertainty.	CGAN pair	Feature, treatment and outcome vectors
Yoon et al.	RadialGAN (RadialGAN)	Multi-domain translation, features and distribution mismatch, cycle-consistency, augmentation.	CGAN, WGAN,	Tabular, discrete and continuous features.
Yoon et al.	Generative Adversarial Imputation Network (GAIN)	Tabular data imputation.	Missing Completly at Random (MCaR) CGAN	Incomplete dataset of continuous variables with entries MCaR.
2019				
Wang et al.	Sequentially Coupled GAN (SC-GAN)	Capturing mutual influence in time-series. Coupled generator pair. Treatment recommendation task.	LSTM CGAN	Continuous patient centric data, including patient state and medication dosage data.
Baowaly et al.	Boundary-seeking medGAN (MedBGAN)	Improving training process.	medGAN Boundary-seeking GAN	Binary counts and occurrences of medical concepts.
Baowaly et al.	MedBGAN, Wassertein medGAN (MedWGAN)	Improving training process.	medGAN BGAN WGAN	Binary counts and occurrences of medical concepts.
Severo et al.	Conditional WGAN-GP (cWGAN-GP)	Generation and public release of dataset. Protecting commercial sensitive information. Class imbalance.	cWGAN-GP, CGAN	Continuous features of vital signs.
Chin-Cheong et al.	WGAN	Heterogeneous mixture of dense and sparse features. Privacy and evaluating the introduction of bias.	WGAN, WGAN-GP, Mode-specific normalization (MSN), DP aware optimizer from Tensor-flow community.	Binary, continuous and categorical features.
Jordon et al.	Private Aggregation of Teacher Ensembles (PATE) framework applied to GANs (PATE-GAN)	Alternative differential privacy, adaptation of the Private Aggregation of Teacher Ensembles (PATE) framework.	Demographic and binary features.	
Torfi and Beyki	corGAN (corGAN)	Convolutional NN (CNN) architecture, capturing feature correlations, evaluating realism, privacy evaluation using Membership Inference (MI).	1D-Convolutional AE (CAE)	Discrete counts and occurrences of medical codes.
Chu et al.	Adversarial Deep Treatment Effect Prediction (ADTEP)	ITE, two independent AE for patient and treatment feature sets, trained adversarially in combination, and outcome predictor from latent representation.	EHR data, not specified	
Jackson and Lussetti	medGAN	Evaluating medgan with the addition of demographics features.	One-hot encoded demographic features and original data.	
Yu et al.	SSL-GAN	Rare disease detection, Semi-supervised learning (SSL), leveraging unlabeled EHR data, medical code embedding network.	LSTM	Diagnosis and prescription codes.
Yang et al.	CGAN	Class imbalance, low count of minority class. Semi-supervised learning combining Self-training (ST) and CT with a CGAN for a IoT application.	Twenty medical datasets from the UCI repository, types unspecified. RFID IoT data of cerebral stroke patients.	
Yang et al.	CorrNN and T-wGAN (GcGAN)	Capturing the correlations between different categories of medical codes and the outcome.	Correlation NN Turing GAN Wassertein T-GAN (T-wGAN)	Binary occurences of medical codes, prescriptions, adjuvants, diagnosis and outcome.
Yang et al.	Categorical GAIN (CGAIN)	Improve on GAIN for categorical variable using fuzzy encoding of the features.		Categorical (multi-class and multi-label) and continuous.

Table 4: Summary of the publication included in the review (Continued).

Publication	Algorithm(s)	Focus and topics	Other algorithms and topics	Data type
Beaulieu-Jones et al.	Auxiliary Classifier GAN (AC-GAN)	Evaluating if differentially private GANs that is valid reanalysis while ensuring privacy.	DP CGAN	Real-valued physiological time-series.
Xu et al.	Conditional Tabular GAN (CTGAN)	Non-Gaussian multi-modal distribution of continuous columns and imbalanced discrete column in tabular data. Evaluation benchmark.	CGAN Training by sampling (TbS) MSN WGAN-GP Gumbel-GAN	Tabular data, mixture of continuous and categorical features.
Yale et al.	HealthGAN	Privacy metrics and over-fitting.	MI, Nearest-neighbor Adversarial Accuracy (NN-AA), Privacy loss (PL), Discriminator testing (DT)	Categorical demographics, continuous vital signs and binary medical codes.
Fisher et al.	Adversarially trained CRMB	Simulation of patient trajectories from their baseline state, disease prediction and risk quantification, missingness.	CRMB	Binary, ordinal, categorical, and continuous, recorded in intervals of 3 months.
2020				
Walsh et al.	Adversarially trained CRMB	Digital twins, disease prediction and risk quantification, missingness.	CRMB	Binary, ordinal, categorical, and continuous, recorded in intervals of 3 months.
Yale et al.	HealthGAN	Metrics to capture a synthetic dataset's resemblance, privacy, utility and footprint. Evaluating applications. Application case studies, Reproducibility of studies with SD.	NN-AA, PL, Data obfuscation (DO, medGAN, WGAN-GP, Synthetic Data Vault,	Continuous and categorical data, demographics, vital signs, diagnoses, and procedures.
Tantipongpipat et al.	DP-auto-GAN (DP-auto-GAN)	Privacy, medGAN adaptation, evaluation metrics.	DP-SGD AE medGAN Renyi Differential Privacy (RDP)	Medical data: binary. Non-health data: categorical and real-valued.
Bae et al.	GANs for anonymizing private medical data (AnomiGAN)	Probabilistic scheme that ensures *indistinguishability* of the SD, than can be viewed as encrypted.	DP CNN	Binary occurences of medical codes.
Cui et al.	*C*omplementary patter**n** **A**augmentatio**n** (CONAN)	Complementary GAN in a rare disease predictor model that generates positive samples from negatives to alleviate class imbalance.	Embedding vectors representing multiple patient visits and conditions.	
Zhu et al.	Blood Glucose GAN (GluGAN)	Adversarialy trained RNN to predict the upcoming time-step in physiological time-series conditioned on the past observations.	RNN, CNN, Gated Recurrent Unit (GRU)	Time-series of blood glucose measurements along with discrete patient submitted features.
Chen et al.	medGAN, WGAN-GP, DC-GAN	Privacy analysis of generative models.	MI Full Black-box Attack, Partial Black-box Attack, White-box Attack, DP-SGD.	Binary feature vector of medical codes.
Chin-Cheong et al.	WGAN with DP (WGAN-DP)	Heterogeneous data, effect of differential privacy on utility.	WGAN DP	Categorical, continuous, ordinal, and binary. Dense or sparse.
Zhang et al.	EMR Wassertein GAN (EMR-WGAN)	Improving training, evaluation metrics, sparsity.	WGAN, BN, Layer normalisation (LN), CGAN	Binary vector of occurrence over the medical codes. Low-prevalence of codes.
Yan et al.	Heterogeneous GAN (HGAN)	Improvements on EMR-WGAN incorporating record-level constraints in the loss function.	WGAN, BN, LN, CGAN, MI, PD	Binary, categorical and continuous.
Ozyigit et al.	Realistic Synthetic Dataset Generation Method (RSDGM)	Exploring the feasibility of various methods to generate synthetic datasets.	WGAN	Continuous and categorical
Yoon et al.	Anonymization through data synthesis using GAN (ADS-GAN)	Identifiability view of privacy. Generator conditioned on real samples inputs with an identifiability loss to satisfy the identifiability constraint.	WGAN WGAN-GP DP alternative	Continuous and binary features.
Goncalves et al.	MC-medGAN	Comparison of GANs with statistical models to generate synthetic data, evaluation metrics.	MI, AD	Categorical and continuous features

3.3 Data Types and Feature Engineering

No publications made use of OHD in its initial form, patient records in EHR composed of many related tables (normalized form). The complexity of a model would grow rapidly when maintaining referential integrity and statistics between multiple tables. The hierarchy by witch these would interact with each other conditionally is no less complicated (Buda et al., 2015; Patki et al., 2016; Zhang and Tay, 2015; Tay et al., 2013). There are however published GAN algorithms made to consume normalized database in their original form. n regards to OHD, feature engineering was used to adapt the data to task requirements, or to a promising algorithms that fit the date characteristics. The data is transformed into one of four modalities: time series, point-processes, ordered sequences or aggregates described in Fig. 5.

Table 5: Types of observational health data and features engineering

Type	Values and structure	Challenges	Features engineering
Time-series *Continuous* *Regular* *Sporadic*	- Timestamped observations - Continuous, ordinal, categorical and/or multi-categorical - Recorded continuously by medical devices, following a schedule by medical professional, or when necessary	- Observations are often MaR across time end dimensions, erroneous, or completely absent for certain patients. - Time-series of different concepts are often highly correlated and their influence on one another must be accounted for.	Imputation coupled with training Regular Data imputation Binning in into fixed-size intervals Combination of binning and imputation
Point-processes	- Series of timestamped observations of one variable or medical concept per patient	-	Series of events reduced to the time interval between each consecutive occurrence.
Ordered sequences	- Ordered vectors representing one or more patients visits - Medical codes associated with the diagnoses, procedures, measurements and interventions	Variable length High-dimensionalLong-tail distribution of codes	Sequences are projected into a trained embedding that preserves semantic meaning according to methods borrowed from NLP
Tabular Denormalized Relational	- Medical and demographic variables aggregated in tabular format - Continuous, ordinal, categorical and/or multi-categorical features	Medical history is aggregated into a fixed-size vector of binary or aggregated counts of occurrences and combined with demographic features.	

3.4 Data oriented GAN development

3.4.1 Auto-encoders and categorical features

In what is to the best of our knowledge, the first attempt at developing a GAN for OHD. Choi et al. focus on the problem posed by the incompatibility of categorical and ordinal features with back-propagation. Their solution is to pretrain an AE to project the samples to and from a continuous latent space representation. The decoder portion is retained along with its trained weights to form a component of medGAN (Choi et al., 2017a). It is incorporated into the generator and maps the randomly sampled input vectors from the real-valued latent space representation back to discrete features. This first exemplar of synthetic OHD generated by GAN inspires a series of enhancements.

Numerous efforts were made to improve the performance of medGAN. Among the first, Camino et al. developed MC-medGAN in which they modified the AE component by splitting its output into a Gumbel-Softmax (Jang et al., 2016) activation layer for each categorical variable and concatenating the results. (Camino et al., 2018). The authors also developed an adaptation based on recent training techniques: WGAN (Arjovsky et al., 2017) and a WGAN with Gradient Penalty (Gulrajani et al., 2017). In brief, the Wasserstein distance is a measure between two Probability Distributions (PDs) that has the property of always providing a smooth gradient. When used as the loss function of the discriminator, it generally improves training stability and mitigates mode collapse. The Wassertein loss function a 1-Lipschitz constraint that was originally solved by weight clipping. It was however demonstrated that in some cases this prevented the network from modelling the optimal function, thus Gradient penalty, a less restrictive regularization was introduced (Petzka et al., 2018). MC-WGAN-GP is the equivalent of MC-medGAN but with Softmax layers. The authors report that the choice of a model will depend on data characteristics, particularly sparsity.

Wasserstein's distance was widely adopted by subsequent authors owing to the propensity of OHD to induce mode collapse. Baowaly et al. developed MedWGAN also based on WGAN, and MedBGAN borrowing from Boundary-seeking GAN (BGAN) (Hjelm et al., 2017) which pushes the generator to

produce samples that lie on the decision boundary of the discriminator, expanding the search space. Both led to improved data quality, in particular MedBGAN (Baowaly et al., 2019; Baowaly et al., 2018). In other effort, Jackson and Lussetti tested medGAN on an extended dataset containing demographic and health system usage information, obtaining results similar to the original (Jackson and Lussetti, 2019). The HealthGAN built upon WGAN-GP, but includes a data transformation method adapted from the Synthetic Data Vault (Patki et al., 2016) to map categorical features to and from the unit numerical range (Yale et al., 2020).

3.4.2 Forgoing the autoencoder and conditional training

Claiming that the use of an AE introduces noise, with EMR-WGAN, Zhang et al. dispose of the AE component of previous algorithms and introduce a conditional training method, along with conditioned BN and LN techniques to stabilise training (Zhang et al., 2020). The algorithm was further adapted by Yan et al. as HGAN to better account for the conditional distributions between multiple data types and enforce record-wise consistency. A recognized problem with medGAN was that it produced commonsense inconsistencies, such as gender mismatches in medical codes (Yan et al., 2020; Choi et al., 2017a). In HGAN, constraints are enforced by adding specific penalties to the loss function, such as limit ranges for numerical categorical pairs and mutual exclusivity for pairs of binary features (Yan et al., 2020). The algorithm also performs well on regular time-series of sleep patterns (Dash et al.)

To develop CTGAN, Xu et al. presume that tabular data poses a challenge to GAN owing to the non-Gaussian multi-modal distribution of continuous columns and imbalanced discrete columns (Xu et al., 2019). Their algorithm, composed of fully connected layers, was developed with adaptations to deal with both continuous and categorical features. For continuous features, it employs mode-specific normalization to capture the multiplicity of modes. For discrete features conditional training-by sampling is devised to re-sample discrete attributes evenly during training, while recovering the real distribution when generating data.

In other efforts, Torfi and Beyki develop corGAN, in which the AE is replaced by a 1-dimensional Convolutional AE (1D-CAE) to capture neighboring feature correlations of the input vectors (Torfi and Beyki, 2019). Chin-Cheong et al. use a Feed-forward Network (FFN) based on Wassertein distance to evaluate the capacity of GANs to model heterogeneous data of dense and sparse medical features (Chin-Cheong et al., 2020). Ozyigit et al. use the same approach with regards to reproducing statistical properties (Ozyigit et al., 2020).

3.4.3 Time-series

Reproducing physiological time-series Esteban et al. devise the RGAN and RC-GAN based on LSTM to generate a regular time-series of physiological measurements from bedside monitors (Esteban et al., 2017). Curiously, the authors dismiss Wassertein's distance, stating that they did not find application in their experiments. In addition, each dimension of their time-series is generated independently from the others, where one would assume they are correlated. A considerable loss of accuracy is observed on their utility metric. In a benchmark on non-health data with long-term dependencies and complex multidimensional relationships against DopelGANger it was outperformed (Lin et al., 2019).

3.5 Task oriented GAN development

3.5.1 Semi-supervised learning

To develop ehrGAN, an algorithm for sequences of medical codes that learns a transitional distribution, Che et al. combine an Encoder-Decoder CNN (Rankin et al., 2020) with VCD (Che et al., 2017). The ehrGAN generator is trained to decode a random vector mixed with the latent space representation of a real patient. The trained ehrGAN model is then incorporated into the loss function of a predictor where it can help generalization by producing neighbors for each input sample.

SSL is commonly employed to augment the minority class in imbalanced datasets, techniques such as ST and CT. Yang et al. improve on both of these by incorporating a GAN in the procedure (Yang et al., 2018). The GAN is first trained on the labelled set and used to re-balance it. A prediction task

with a classifier ensemble is then executed and the data points with highest prediction confidence are labelled. The process is iterated until labelling expansion ceases. As a final step, the GAN is trained on the expanded labelled set to generate an equal amount of augmentation data. The authors obtained improved performance in a number of classification tasks and multiple tabular datasets with their method.

3.5.2 Domain translation

To address the heterogeneity of healthcare data originating from different sources, Yoon et al. combines the concepts of cycle-consistent domain translation from Cycle-GAN (Zhu et al., 2017b) and multi-domain translation from Star-GAN (Choi et al., 2017b) to build RadialGAN to translate heterogeneous patient information from different hospitals, correcting features and distribution mismatches (Yoon et al., 2018c). The algorithm use an encoder-decoder pair per data endpoint that are trained to map records to and from a shared latent representation for their respective endpoint.

3.5.3 Individualized treatment effects

The task of estimating ITEs is an ongoing problem. ITEs refer to the response of a patient to a certain treatment given a set of characterizing features. This is due to the fact that counterfactual outcomes are never observed or treatment selection is highly biased (Yoon et al., 2018a; McDermott et al., 2018; Walsh et al., 2020). To overcome this problem, in GANITE Yoon et al. employ a pair of GANs, one for counterfactual imputation and another for ITE estimation (Yoon et al., 2018a). The former captures the uncertainty in unobserved outcomes by generating a variety of counterfactuals. The output is fed to the latter, which estimates treatment effects and provides confidence intervals.

With CWR-GAN, a joint regression-adversarial model, McDermott et al. demonstrated a SSL approach inspired by Cycle-GAN to leverage large amounts of unpaired pre/post-treatment time-series in Intensive Care Unit (ICU) data for the estimation of ITEs on physiological time-series (McDermott et al., 2018). The algorithm has the ability to learn from unpaired samples, with very few paired samples, to reversibly translate the pre/post-treatment physiological series.

Chu et al. approach the problem of data scarcity by designing ADTEP, an algorithm that can maximize use of the large volume of EHR data formed by triples of non-task specific patient features, treatment interventions and treatment outcomes (Chu et al., 2019). ADTEP learns representation and discriminatory features of the patient, and treatment data by training an AE for each pair of features. In addition to AE reconstruction loss, a second model is tasked with advsersarially identifying fake treatment feature reconstructions. Finally, a fourth loss metric is calculated by feeding the concatenated latent representations of both AEs to a Logistic-regression (LR) model aimed at predicting the treatment outcome (Chu et al., 2019).

Similarly to Esteban et al., Wang et al. demonstrated an algorithm to generate a time series of patient states and medication dosages pairs using LSTM. In contrast to RGAN and RC-GAN, in SC-GAN, patients state at the current time-step informs the concurrent medication dosage, which in turn affects the patient state in the upcoming time-step (Wang et al., 2019). SC-GAN overcame a number of baselines on both statistical and utility metrics.

3.5.4 Data imputation and augmentation

GAN are naturally suited for data imputation, and could provide a new approach to deal with the problems of health data relating to widespread missingness. Statistical models developed for the multiple imputation problem increase quadratically in complexity with the number of features, while the expressiveness of deep neural networks can efficiently model all features with missing values simultaneously.In that regard, Yoon et al. adapted the standard GAN to perform imputation on continuous features MaR in tabular datasets (Yoon et al., 2018b). In GAIN, the discriminator is tasked with classifying individual variables as real or fake (imputed), as opposed to the whole ensemble. Additional input, or hint, containing the probability of each component being real or imputed is fed to the discriminator to resolve the multiplicity of optimal distributions that the generator could reproduce. The model

performs considerably better than five state-of-the-art benchmarks. GAIN was later adapted to also handle categorical features using fuzzy binary encoding, the same technique employed in HealthGAN.

The distribution estimated by a generator model can compensate for lack of diversity in a real sample, essentially filling in the blanks in a manner comparable to data imputation. In such cases, data sampled from this distribution has the potential to help improve generalization in training predictive models. We find evidence of this in generating unobserved counterfactual outcomes (Yoon et al., 2018b), or generating neighboring samples to help generalization in predictors (Che et al., 2017). The adversarially trained Restricted Boltzmann Machine (RMB) developed by Fisher et al. enabled them to simulate individualized patient trajectories based on their base state characteristics. Due to the stochastic nature of the algorithm, generating a large number of trajectories for a single patient can provide new insights on the influence of starting conditions on disease progression or quantify risk (Fisher et al., 2019).

3.6 Model validation and data evaluation

To asses the solution to a generative modelling problem, it is necessary to validate the model, and to verify its output. GAN aim to approximate a data distribution P, using a parameterized model distribution Q (Borji, 2019). Thus, in evaluating the model, the goal is to validate that the learning process has led to a sufficiently close approximation. What this means in practice is hard to define. The concept of "realism" finds more natural application to images of text, but is more ambiguous when faced with the complexity of health data. Walsh et al. employ the term "statistical indistinguishability" and define it as the inability of a classification algorithm to differentiate real from synthetic samples (Walsh et al., 2020). The terms covers almost all evaluation methods employed in the publications, which can be divided into two broad categories: those aimed at evaluating the statistical properties of the data directly, and those aimed at doing so indirectly by quantifying the work that can be done with the data. There are, nonetheless a few attempts of a qualitative nature, more in line with the concept of realism.

3.6.1 Qualitative evaluation

Visual inspection of projections of the SD is a common theme, serving mostly as a basic sanity check, but occasionally presented as evidence. The formal qualitative evaluation approaches found in the literature are mainly Preference Judgement, Discrimination Tasks or Clinician Evaluation and are generally carried out by medical professionals in the appropriate field (Borji 2018).

- **Preference judgment** The task is choosing the most realistic of two data points in pairs of one real and one synthetic (Choi et al., 2017a).
- **Discrimination Tasks** Data points are shown one by one and must be classified as real or synthetic (Beaulieu-Jones et al., 2019).
- **Clinician Evaluation** Rather than classifying the data points, they must be rated for realism according to a predefined numerical scale. (Beaulieu-Jones et al., 2019). Significance is determined with a statistical test such as Mann-Whitney.
- **Visualized embeding** The real and synthetic data samples are plotted on a graph or projected into an embeding such as t-SNE or PCA and compared visually. (Cui et al., 2019; Yu et al., 2019; Zhu et al., 2020; Yale et al., 2019a; Yang et al., 2019c; Beaulieu-Jones et al., 2019; Tantipongpipat et al., 2019; Dash et al.).
- **Feature analysis** In certain fields, the data can be projected to representations that highlight patterns or properties that can be easily visually assessed. While this does not provide conclusive evidence of data realism, it can help get a better understanding of model behaviour during training. As an example, typical and easily distinguishable patterns in EEG and ECG bio-signals. (Harada et al., 2019)

In general, qualitative evaluation methods based on visual inspection are weak indicators of data qual-

ity. At the dataset or sample level, quantitative metrics provide more convincing evidence of data quality (Borji 2018).

3.6.2 Quantitative evaluation

Quantitative evaluation metrics can be categorized into three loosely defined groups: those comparing the distributions of real and synthetic data as a whole, those aimed at assessing the marginal and conditional distributions of features, and those evaluating the quality of the data indirectly by quantifying the amount of work that can be done with the data, referred to as utility.

- **Dataset distributions** A summary of metrics based on comparing distributions is presented in Tab. 6.
- **Feature Distributions** If the model has learned a realistic representation of the real data it should produce SD that possesses the same quantity and type of information content. Authors attempt by various metrics to determine if the statistical properties of the SD agree with those of the real data. These metrics are presented in Table 7. Although statistical similarity provides strong support for the behavior of the learning process, it is not necessarily informative about their validity. They are often ambiguous and can be found to be misleading upon further investigation. Given the complexity of health data, low level relations are unlikely to paint a full picture. Authors often state that no single metric taken on its own was sufficient, and that a combination of them allowed deeper understanding of the data.
- **Data utility** Utility-based metrics often provide a more convincing indicator of data realism, on the other hand they mostly lack the interpretability that some statistical metrics allow. These are presented in Table 8. We took the liberty of placing these into one of two categories: tasks mostly defined for evaluation (Ad hoc utility metrics) or tasks based on real-world applications (Application utility metrics). Note that this distinction is not based on a rigorous definition, but serves to facilitate understanding.
- **Analytical** The analytical methods were mainly employed for evaluation, but can also provide a better understanding of the and its behavior.
  - *Feature Importance* The important features (Random Forest (RF)) and model coefficients (LR, Support Vector Machine (SVM)) of predictors trained for some task are compared between real and synthetic data. (Esteban et al., 2017; Xu et al., 2019; Yoon et al., 2020; Chin-Cheong et al., 2019; Beaulieu-Jones et al., 2019).
  - *Ablation study* The performance of the model is compared against versions impared version, with some components removed. This helps determining if the novel component of the algorithm contributes significantly to performance (Cui et al., 2019; Che et al., 2017; McDermott et al., 2018; Yoon et al., 2018c; Chin-Cheong et al., 2020).

Table 6: Metrics employed to validate trained models based on the comparison of distributions.

Metric	Description
Kullback-Leibler divergence (KLD)	Non-symmetric measure of difference between two PDs, related to relative entropy. Given a feature X, $p(x)$ and $q(x)$ the PD of the real and synthetic data respectively, the KLD of $q(x)$ from $p(x)$ is the amount of information lost when $q(x)$ is trained to estimate $p(x)$ (Jiawei, 2018; Goncalves et al., 2020).
RDP	Alternative measure of divergence, which includes KLD as a special case. The RDP includes a parameter α that gives it an extra degree of freedom, becoming equivalent to the Shannon-Jensen divergence when $\alpha \longrightarrow 1$. It showed a number of advantages when compared to the original GAN loss function, and removed the need for gradient penalty (Van Balveren et al., 2018; Tantipongpipat et al., 2019)
Jaccard similarity	Measure of similarity and diversity defined on sets as the size of the intersection over the size of the union (Ozyigit et al., 2020; Yang et al., 2019c; contributors).
2-sample test (2-ST)	Statistical test of the null hypotheses the real and SD samples came from the same distribution. and synthetic, originate from the same distribution through the use of a statistical test such as Kolmogorov-Smirnov (KS) or Maximum Mean Discrepency (MMD).(Fisher et al., 2019; Baowaly et al., 2019; Baowaly et al., 2018; Esteban et al., 2017)
Distribution of Reconstruction Error	Compares the distributions of reconstruction error for the SD and the training set versus the SD and a held out testing set. Calculated according to the Nearest-neighbor metric or other measures of distance. A significant difference would indicate over-fitting and can evaluated with a statistical test, such as KS. (Esteban et al., 2017)
Latent space projections	Real and synthetic samples are projected back into the latent space, or encoded with a β variational auto-encoder (β-VAE), comparing the dimensional mean of the variance or the distance between mode peaks (Zhang et al., 2020). See Section 6.2.1 for examples of how the latent space encoding can interpreted.
Domain Specific Measures (DSMs)	Comparison of the PD with DSMs. For instance the Quantile-Quantile (Q-Q) plot for point-processes (Xiao et al., 2017). See Section 6.2 for a notion of how DSMs could apply to EHR data.
Classifier accuracy	Accuracy of a classifier trained to discriminate real from synthetic units. Predictor accuracy around 0.5 would indicate indistinguishability. (Fisher et al., 2019; Walsh et al., 2020)

Table 7: Metrics based on evaluating the statistical properties of the synthetic data distribution.

Metric	Description
Dimensions-wise distribution	The real and synthetic data are compared feature-wise according to a variety of methods For example, the Bernoulli success probability for binary features, or the Student T-test for continuous variables, and Pearson Chi-square test for binary variables is used to determine statistical significance (Beaulieu-Jones et al., 2019; Choi et al., 2017a; Chin-Cheong et al., 2019; Yan et al., 2020; Baowaly et al., 2019; Baowaly et al., 2018; Ozyigit et al., 2020; Tantipongpipat et al., 2019; Yoon et al., 2020; Tantipongpipat et al., 2019; Fisher et al., 2019; Che et al., 2017; Wang et al., 2019; Yale et al., 2019a; Chin-Cheong et al., 2020; Ozyigit et al., 2020).
Inter-dimensional correlation	Dimension-wise Pearson coefficient correlation matrices for both real and synthetic data (Beaulieu-Jones et al., 2019; Goncalves et al., 2020; Torfi and Beyki, 2019; Frid-Adar et al., 2018; Ozyigit et al., 2020; Yang et al., 2019c; Yoon et al., 2020; Zhu et al., 2020; Yoon et al., 2020; Walsh et al., 2020; Yale et al., 2019a; Ozyigit et al., 2020; Dash et al.; Bae et al., 2020b).
Cross-type Conditional Distribution	Correlations between categorical and continuous features, comparing the mean and standard deviation of each conditional distribution (Yan et al., 2020).
Time-lagged correlations	Measures the correlation between features over time intervals. (Fisher et al., 2019; Walsh et al., 2020).
Pairwise mutual information	Checks for the presence multivariate relationships pair-wise for each feature, as a measure of mutual dependence (Rankin et al., 2020). Quantifies the amount of information obtained about a feature from observing another.
First-order proximity metric	Defined over graphs, captures the direct neighbor relationships of vertices. Zhang et al. applied to graphs built from the co-occurrence of medical codes and compared the results between real and synthetic data (Zhang et al., 2020).
Log-cluster metric	Clustering is applied to the real and synthetic data combined. The metric is calculated from the number of real and synthetic samples that fall in the same clusters (Goncalves et al., 2020).
Support coverage metric	Measures how much of the variables support in the real data is covered in the synthetic data. Support is defined as the percentage of values found in the synthetic data, while coverage is the reverse operation. The metric is calculated as the average of the ratios over all features. Penalizes less frequent categories that are underrepresented (Goncalves et al., 2020).
Proportion of valid samples	Defined by Yang et al. as a requirement for records to contain both disease and medication instances. (Yang et al., 2019c).
PCA Distributional Wassertein distance	The Wassertein distance is calculated over k-dimensional PCA projections of the real and synthetic data (Tantipongpipat et al., 2019).

Table 8: Metrics based on evaluating the utility of the synthetic data on practical tasks.

Metric	Description
	Data utility metrics
DWP	Each variable is in turn chosen as the prediction target label and the remaining as features. Two predictors are trained to predict the label, one from the synthetic data and another from a portion of the real data. Their performance is compared on the left out real data (Choi et al., 2017a; Camino et al., 2018; Goncalves et al., 2020; Yan et al., 2020; Tantipongpipat et al., 2019; Baowaly et al., 2019).
ARM	ARM aims to the discovery of relationships among a large set of variables, commonly occurring variable-value pairs (Agrawal et al., 1993). The rules obtained from the real and synthetic data are compared (Baowaly et al., 2019; Baowaly et al., 2018; Bae et al., 2020a; Yan et al., 2020).
Training utility	Performance of predictors trained on the synthetic data, often in comparison with the real data or data generated with DP (Bae et al., 2020a).
TRTS	Accuracy on real data of some form of predictor trained on synthetic data (Beaulieu-Jones et al., 2019; Rankin et al., 2020; Yoon et al., 2020).
TSTR	Accuracy on synthetic data of some form of predictor trained on real data (Bae et al., 2020a; Yoon et al., 2020; Jordon et al., 2019).
Discriminator	A predictor is trained to discriminate synthetic from real sample. An accuracy value of 0.5 would indicate that they are indistinguishable (Fisher et al., 2019; Walsh et al., 2020; Yale et al., 2019b).
Siamese discriminator	A pair of identical FFN each receive either a real sample or a synthetic sample. Their output is passed to a third network which outputs a measure of similarity (Torfi and Beyki, 2019).
	Applied utility metrics
Data augmentation	A predictor is trained on a combination dataset of real and synthetic data or real data with missing values imputed and performance is compared with the same predictor trained on real data alone (Yoon et al., 2020; Yang et al., 2019b,c).
Model augmentation	The trained generative model is incorporated into a predictor's activation function by generating an ensemble of proximate data points for each instance, thereby improving generalization (Che et al., 2017).
Accuracy	The prediction performance of the model is compared against benchmarks of the same type on real data (Cui et al., 2019; Yoon et al., 2018a; Che et al., 2017; Yu et al., 2019; Zhu et al., 2020; Baowaly et al., 2019; Wang et al., 2019; Walsh et al., 2020; Yoon et al., 2018b; McDermott et al., 2018; Yang et al., 2019c; Yoon et al., 2018c; Xu et al., 2019; Beaulieu-Jones et al., 2019; Bae et al., 2020a). Models trained to make forward predictions from past observations or from real data transformed with a known function can simply be evaluated for accuracy. For example, the RMSE on time-series (Xiao et al., 2018b; McDermott et al., 2018; Yoon et al., 2018b; Yang et al., 2019b; Zhu et al., 2020).

3.7 Alternative evaluation

In their publications, Yale et al. propose refreshing approaches to evaluating the utility of SD. For example, they organized a hack-a-thon type challenge involving the data. During the event, students were tasked with creating classifiers, while provided only with SD (Yale et al., 2020). They were then scored on the accuracy of their model on real data. In more rigorous initiatives, they attempted (successfully) to recreate the experiments published in medical papers based on the MIMIC dataset using only data generated from their model HealthGAN. In a subsequent version of their article, the authors evaluate the performance of their model against traditional privacy preservation methods by using the trained discriminator component of HealthGAN to discriminate real from synthetic samples.

3.8 Privacy

Some authors offered a privacy risk assessment of their SD To evaluate the risk of reidentification, empirical analyses were conducted according to the definitions of MI, AD (Choi et al., 2017a; Goncalves et al., 2020; Yan et al., 2020; Chen et al., 2019b; Chin-Cheong et al., 2020) and the Reproduction rate (RR) (Zhang et al., 2020). Cosine similarities between pairs of samples are also employed (Torfi and

Beyki, 2019). Most studies report low success rates for these types of attacks, and little effect from the sample size, although Chen et al. note that sample sizes under 10k lead to higher risk.

Some have put forward the notion that preventing over-fitting and preserving privacy may not be conflicting goals (Wu et al., 2019; Mukherjee et al., 2019). Numerous attempts have been made to apply traditional privacy guarantees, such as deferentially-private stochastic gradient descent (Beaulieu-Jones et al., 2019; Esteban et al., 2017; Chin-Cheong et al., 2020; Bae et al., 2020a). By limiting the gradient amplitude at each step and adding random noise, AC-GAN could produce useful data with $\epsilon = 3.5$ and $\delta < 10^{-5}$ according to the definition of differential privacy. Uniquely, Bae et al. ensure privacy with a probabilistic scheme that ensure indistinguishably, but also maximizes utility. Specifically, a multiplicative perturbation by random orthogonal matrices with input entries of kxm medical records and a second second discriminator in the form of a pretrained predictor (Bae et al., 2020a). In black-box and white-box type attacks, including the LOGAN (Hayes et al., 2017) method, medGAN performed considerably better than WGAN-GP (Chen et al., 2019b), the algorithm which served as basis for improvements to medGAN in publications discussed in Section 3.4.1. Overall, the author notes that releasing the full model poses a high risk of privacy breaches and that smaller training sets (under 10k) also lead to a higher risk. Goncalves et al. evaluated MC-medGAN against multiple non-adversarial generative models in a variety of privacy compromising attacks, including AD, obtaining inconsistent results for MC-medGAN (Goncalves et al., 2020). While this is not mentioned by the authors, multiple results reported in the publication point to the fact that the GAN was not properly trained or suffered mode-collapse.

Means to confer privacy guarantees on SD generated by GAN are being actively researched in a variety of fields, many of which are a priori readily applicable to health data. At this stage, however, contradictory results have between obtained where the statistical fidelity of the synthetic seemed to be preserved, but utility-based measures based on a classification were degraded by incorporating DP. S In privGAN, Mukherjee et al., an adversary is introduced, forcing the generator to produce samples that minimize the risk of MIA attack, in addition to cheating the discriminator. The combination of both goals has the explicit effect of preventing over-fitting, and their algorithm produces samples of similar quality to non-private GAN.

3.8.1 The status of fully synthetic data in regards to current privacy regulations

It seems intuitively possible that the artificial nature of SD essentially prevents associations with real patients, however the question is never directly addressed in the publications. An extensive Stanford Technological Review legal analysis of SD concluded that laws and regulations should not treat SD indiscriminately from traditional privacy preservation methods (Bellovin et al., 2019). They state that current privacy statutes either outweigh or downplay the potential for SD to leak secrets by implicitly including it as the equivalent of anonymization.

3.8.2 Alternative views of privacy

The discordance between the theoretical concepts of DP, which are based ultimately on infinite samples, and the often insufficient data on which the probability of disclosure is calculated remains deficient. Therefore, Yoon et al. have postulated an intriguing alternative view of privacy (Yoon et al., 2020). They propose to emphasize measuring identifiability of finite patient data, rather than the probabilistic disclosure loss of DP based on unrealistic premises. Simplistically, they define identifiability as the minimum closest distance between any pair of synthetic and real samples. In their implementation, the generator receives both the usual random seed and a real sample as input. This has the effect of mitigating mode collapse, but also of reproducing the real samples. On the other hand, the discriminator is equipped with an additional loss metric based on a measure of similarity between the original sample and the generated one, thus ensuring the tuneable threshold of identifiability is met. Their results on a number of previously discussed evaluation metrics are encouraging.

In a similar approach, Yale et al. broke away from the theoretical guarantees of traditional methods with a measure native to GAN. Their proposal is a metric quantifying the loss of privacy, a concept

more aligned with the objective of GAN to minimize the loss of data utility (Yale et al., 2019b,c). They point out, quite appropriately, the advantage of concrete measurable values of loss in utility and privacy when making the decision of releasing sensitive data. Briefly, the Nearest Neighbor Adversarial Accuracy measures the loss in privacy based on the difference between two nearest neighbor metrics. The first component is the proportion of synthetic samples that are closer to any real sample than any pair of real samples. The second component is the reverse operation. In a subsequent paper, HealthGAN evaluated against traditional privacy preservation methods with a variant of the IA based on the nearest neighbor metric. HealthGAN performs considerably better than all other methods, while still maintaining utility on a prediction task.

4 Discussion

4.1 Applications of GANs for health data and innovation

Overall, the published GAN algorithms for OHD provided equivalent or superior performance against the statistical modeling-based methods against which they were benchmarked. Importantly, their capabilities are highly relevant to the medical field: domain translation for unlabeled data, conditional sampling of minority classes, data augmentation, learning from partially labeled or unlabeled data, data imputation, and forward simulation of patient profiles. While some of these claims are overoptimistic or lack convincing evidence, they paint an encouraging picture for the value of synthetic OHD and the transformative effect it could have on healthcare initiatives and scientific progress.

4.2 Challenges posed by OHD

The challenges posed by health data are obvious, and a number of recurrent factors influenced the outcome of efforts to develop GANs for OHD. These problems are not limited to generative algorithms, but also ML in general. While the progress in developing new algorithms has great momentum, their application and adoption will undoubtedly be more sluggish, as has been the case with predictive ML.

In the case of generative models, multi-modality is one aspect that caused the most trouble in achieving a stable training procedure. At the outset, preventing mode collapse was an issue that attracted the most research efforts, in addition to data containing combinations of categorical and continuous features. A rapid succession of efforts aimed at improving medGAN by incorporating the latest machine learning techniques, known to improve performance across a broad range of applications, showed continued improvements. However, taken as a whole the efforts were haphazard and often yielded unsurprising results. This is not unexpected in a new field, and more concerted efforts to systematically approach the problems would surely formalize the research.

While the problem of mode collapse has been alleviated, evidence has yet to be provided with regards to ensuring that the finer details of the distribution are estimated with sufficient granularity to produce realistic patient profiles. In this direction conditional training methods have led to improvements. For example, when labels corresponding to sub-populations or classes are used to condition the generative process. Zhang et al. showed that conditioned training with categorical labels, in this case age ranges, improves utility for small datasets, but not with larger samples (Zhang et al., 2020) As described in Section 3.4.2, HGAN further introduces constraint-based loss. Based on the distribution of individual features and utility-based metrics, the authors argue that the bias intrinsic to their methods has not led to undesirable bias or side-effects in other aspects of the learned distribution.

The idea of introducing human knowledge in the otherwise naive training process has gained some attention. Not only can this improve the speed and quality of training, but also implies some degree of interpretability.

4.3 Evaluation metrics and benchmarking

In regards to the practices of evaluation, the choice of optimal metrics and indicators is still being explored. Overall, no evaluation metric proposed addresses the concept of realism in synthetic data. The blatant observation is that the efforts are far from consistent or systematic. This has led to a number of issues. As a striking example, competing methods are often compared with different metrics or with contradictory results in different datasets (Baowaly et al., 2019; Baowaly et al., 2018; Camino et al., 2018; Choi et al., 2017a; Zhang et al., 2020). In their evaluation of medGAN, Yale et al. argue that the positive resemblance of plotted feature distribution of synthetic data against real data is due to the fact that the model's architecture tends to favor reproducing the means and probabilities of each diagnosis column. For example, synthetic data contains samples with an unusually high number of codes. Their hypothesis is that these samples are used by the algorithm to discharge the rare medical codes with weak correlation to balance the distributions. However, they stated in their experiments that comparing PCA plots of real and synthetic data for various generation methods was insightful to get an impression of their behavior (Yale et al., 2020).

Qualitative evaluation, in its current form, provides little evidence. For medical experts, these representations are meaningless. As such, the results of qualitative evaluation often state that synthetic data is indistinguishable from the real data (Choi et al., 2017a; Wang et al., 2019). It is doubtful that they could in fact be. Esteban et al. found that participants avoided the median score and were not confident enough to choose either extreme (Esteban 2017).

Reproducing aggregate statistical properties is rather unconvincing evidence that a model has learned to reproduce the complexity of patient health trajectories. Choi et al. found that although the synthetic sample seemed statistically sound, it contained gross errors such as gender code mismatches and suggested the use of domain-specific heuristics (Choi et al., 2017a). HGAN was an encouraging step in this direction, but it may be difficult to scale. In some cases the statistical metrics may be contradictory, such as when the ranking of medical frequencies are wrong, but the data augmentation leads to improved performance (Che et al., 2017). Utility-based metrics provide a more solid evaluation of data quality. However, these metrics only confirm the value of the data according to a narrow context. They are indicative of realism so far as a patient's state is indicative of a medical outcome. Moreover, they do not provide any insight about the validity of the relations found in a patient record and its overall consistency.

In this regard the replication of medical studies with synthetic data by Yale et al. substantiate the value of SD for exploratory data analysis, reproducibility on restricted data and more generally education in scientific training. Reproducing medical or clinical studies will be necessary to gain mainstream adoption of GAN produced SD and dispel the scepticism it is generally met with. The medical domain is known for its slow pace in adopting new technologies and predictive ML is still far from meeting its full implementation potential.

Data utility

More research efforts should by directed to demonstrating that synthetic data generated by GANs posses sufficient utility for scientific analyses. Reproducing published reslults of medical research is a straightforward and convincing way to acheive this.

4.4 Analysis of OHD-GAN

4.4.1 Data representation and algorithm architecture

We observed that majority of methods included in the review made use of altered representations of patient records. Namely, through feature engineering the data is transformed from its original form. This is in part due to the inconvenient properties of health data, such as missingness. However, it is somewhat apparent that the main motive is to accommodate existing algorithms. Along with

demographic variables, OHD data mostly takes the form of triples composed by a timestamp, a medical concept and the recorded value. Their count is different for each patient, irregular intervals between each triple and the number of possible values in a dimensions can be huge. Moreover, there are generally multiple episodes of care, each with a different cause. The form and content is not typically considered practical for machine learning.

At varying degrees, depending on the transformations, information is being lost or bias is being introduced. For example, when data are reduced by aggregation to one-hot encoding of binary or count variables, the complex relationships found in medical data are, for the most part, lost. Similarly, information is lost when forcing continuous time-series into a regular representation, by truncating, padding, binning or imputation. Moreover, it is highly unlikely that the data is missing at random, introducing the potential for bias when a large part of the real data is rejected on this basis. Truncating the medical codes to their parent generalizations (Zhang et al., 2020; Choi et al., 2017a). In brief, loss of information content is being preferred by molding and discarding arbitrarily the data to the benefit of performance metrics, as opposed to the more tricky alternative of developing algorithms according the data.

Deep architectures are based on the intuition that multiple layers of nonlinear functions are needed to learn complicated high-level abstractions (Bengio, 2009). CNN capture patterns of an image in a hierarchical fashion, such that in sequence, each layer forms a representation the data at a higher level of abstraction. This type of data-oriented architecture has led to impressive performance for CNN and image data. The same principle can be applied to health data. An algorithm developed in a hierarchical structure, was demonstrated to form representations of EHR that capture the sequential order of visits and co-occurrence of codes in a visit have led to improved predictor performance, and also allowed for meaningful interpretation of the model (Choi et al., 2016). Similarly, models of time-series based on a continuous time representation, such as found in EHR data, have shown improved accuracy over discrete time-representations (Rubanova et al., 2019; De Brouwer et al., 2019). Nonetheless, creative adaptations of the data for existing architectures have provided surprising results. For example, OHD input into a CNN were transformed to image(bitmaps) in which the pixels encoded the information (Fukae et al., 2020).

5 Recommendations

5.1 Basic models

Overall, evaluation methods were superficial or uni-dimensional. Finding convincing and robust evaluation metrics for synthetic health data is an open issue. Even more so when the learning task is poorly defined or the scope of the problem is too large. The difficulty of explaining or validating the realism of data representing a patient, often longitudinal and which factors deferentially contribute to disease characterization makes the assessment of synthetic data ambiguous, thus demanding stronger evidence to claims.

Modelling efforts for OHD-GAN should be limited in scope to a single data type or modality. This is favourable for a number of evaluation related aspects. Firstly, it makes qualitative evaluation by visual inspection from experts possible and meaningful. Secondly, for same reasons, the behaviour of the model can be assessed straightforwardly. The generative process can be influenced intentionally to observe the effect on the properties of the output. Finally, it allows for quantitative evaluation with domain specific metrics. The scope should clearly identify the purpose of the data generation, its utility and the target patients(Capobianco, 2020; Kappen et al., 2016; Kappen and Peelen, 2016)

5.2 Data-driven architecture

The algorithm architecture of OHD-GAN should be engineered to match the process that generated the data, not the other way around. Data should be used and generated in the form it is first collected.

In addition to preventing information loss, this ensures models will reflect the real generative process. Such models are more likely to provide insights into the system they are taught to imitate and further our understanding about them. Furthermore, the learned statistical distribution is inevitably more meaningful and interpretable, facilitating applications in the healthcare domain and supporting the inference of insights from the learned model parameters.

5.3 Interpretability

Even though a few authors explored the behavior of their models according to various methods, the subject was left largely unmentioned. It is imperative that future experimentation and publication give equal importance to evaluating the interpretation of their models and means to do so, as for performance. In the healthcare domain, black box machine learning models find little adoption, and synthetic data is most often met with attacks to its validity.

6 Directions for future research

6.1 Building a patient model

The ultimate goal for generative models of OHD must be to develop an algorithm capable of learning an all encompassing patient model. It would then be possible to generate full EHR records on demand, integrating genetic, lifestyle, environmental, biochemical, imaging, clinical information into high-resolution patient profiles (Capobianco, 2020). This is in fact the intention of the patient simulator Synthea. However, Synthea will eventually face a problem with scalability and the capacity of semi-independent state-transition models to coordinate in capturing long-range correlations.

Once basic models of health data, as described in Section 5.1, have been developed and validated, these can be progressively combined in a modular fashion to obtain increasingly complex patient simulators. Furthermore, having designed the architecture of these basic models on the underlying data in a way that is comprehensible, as described in 5.2, will facilitate the composition of more complex models. Inputs, outputs and parts of these models can be conditionally attached to others such that the generative process occurs in a way that reflects the real generative process.

6.2 Evaluating complex patient models

Once more complex models are developed, the problem is again finding meaningful evaluation metrics of data realism. Capobiano et al. insist on the necessity for data performance metrics encompassing diagnostic accuracy, early intervention, targeted treatment and drug efficacy (Capobianco, 2020). In their publication exploring the validation of the data produced by Synthea, Chen et al. provide an interesting idea to achieve this (Chen et al., 2019a). Noting that the quality of care is the prime objective of a functional healthcare system, they suggest using Clinical Quality Measures (CQMs) to evaluate the synthetic data. These measures "are evidence-based metrics to quantify the processes and outcomes of healthcare", such as "the level of effectiveness, safety and timeliness of the services that a healthcare provider or organization offers."(Chen 2019). High-level indicators such as CQMs domain specific measures of quality, specifically designed for higher level or multi-modal representations of healthcare data. The constraints introduced in HGAN should be leverage to evaluate the realism of the synthetic data, rather than bias the generator training. Composing a comprehensive set of such constraints could possibly serve as a standardized benchmark. At the individual level, Walsh et al. employ domain specific indicators of disease progression and worsening and compare agreement of the simulated patient trajectories with the factual timelines (Walsh et al., 2020).

In addition to CQM, we propose the use of the Care maps used by the Synthea model to simulate patient trajectories as evaluation metrics (Walonoski et al., 2017). Care maps are transition graphs developed from clinician input and Clinical Practice Guidelines, of which the transition probabilities are gathered from health incidence statistics. While these allow the Synthea algorithm to simulate patient profile with realistic structure, they also prevent it from reproducing real-world variability. Conversely,

while GANs have the ability to reproduce the quirks of real data, they also lack the constraints preventing nonsensical outputs. As such, Care maps provide an ideal metric to check if the synthetic data conforms to medical processes.

In fact, has been used before in a competition where participants were given synthetic data from finite state transition machines with know probabilities and tasked to build and learn models that would reproduce those of the original, unseen models. The participants according to the Perplexity metric. Commonly used in NLP, quantifies how well a probability distribution or probability model predicts a sample (Verwer et al., 2013). We postulate that the Synthea models built with real-world probabilities would provide a unique and robust way to evaluate synthetic data according to the metric proposed above, among other means to utilize the state-transition in Synthea and their modularity.

6.2.1 Latent space

The latent space representation, the lower-dimensional vector space of the data, can provide means for evaluation and interpretability and it's potential should be explored when developing algorithms (Liu et al., 2019). Numerous publications have shown that they capture meaningful properties and structure of the data, reducing complexity to a level that lends itself to interpretation (Way et al., 2020; Koumakis, 2020). In one instance involving transcription factor micro-array data, a close one-to-one mapping could be obtained from the last hidden layer, in addition to the higher level layers that related to biological processes in a hierarchical fashion (Chen et al., 2016). Pushing the boundaries further, by correlating the output features of a GAN with the latent space dimensions allowed controllable semantic manipulation of the generated data (Wang et al., 2020b; Ding et al., 2020; Li et al., 2020), or provided new insights by exploring structured perturbations (Liu et al., 2019).

6.2.2 Opportunities and application to current events

Synthetic and external controls in clinical trials are becoming increasingly popular (Thorlund et al., 2020). Synthetic controls refer to cohorts that have been composed from real observational cohorts or EHR using statistical methodologies. While the individuals included in the cohorts are usually left unchanged, micro-simulations of disease progression at the patient level are used to explore long-term outcomes and help in the estimation of treatment effects (Thorlund et al., 2020; Etzioni et al., 2002). Synthetic data generated by GANs could be transformative for the problem of finding control cohorts.

With the COVID-19 pandemic scientists have become increasingly aware of and vocal about the need for data sharing between political borders (Cosgriff et al., 2020; Becker et al., 2020; McLennan et al., 2020). An obvious application is generating additional amounts of data in the early stages of the pandemic, potentially creating opportunities earlier. Synthetic is data not only an opportunity to facilitate the exchange of data, but also adjust the biases of samples obtained from different localities. Factors such as local hospital practices, different patient populations and equipment introduce feature and distribution mismatches (Ghassemi et al., 2020). These disparities can be mitigated by translation of GAN algorithms, such as Cycle-GAN proposed by Yoon et al.

7 Source-code and datasets

The algorithms presented in this review can undoubtedly find usefulness for other health data or similar problems. Most importantly they can be reevaluated on other datasets or improved by adapting them with latest ML techniques. We present in Table 9 a list of links to the source code published by the authors. In addition, we present in Table 10 the datasets which were employed by the authors in their experiments, for those who were referenced and available. A broad variety of articles about generative and predictive algorithms published along with the source-code can be on Papers With Code in the medical section. Notably, they host a yearly ML Reproducibility Challenge to "[...] encourage the publishing and sharing of scientific results that are reliable and reproducible." in which papers accepted for publication in top conferences are evaluated by members of the community reproducing their experiments (Sinha et al., 2020). Benchmarks are also presented on the website, but unfortunately corGAN

is the only entry in the medical section.

Table 9: Source code and data released and made open-source by the authors

Algorithm	Format	Location	Source code
AC-GAN (Beaulieu-Jones et al., 2019)	Jupyter notebook	GitHub	greenelab/SPRINT_gan
Ward2ICU (Severo et al., 2019)	Python	GitHub	3778/Ward2ICU
AnomiGAN (Bae et al., 2020a)	Python, Tensorflow	Github	hobae/anomigan
GAIN (Yoon et al., 2018b)	Python, Tensorflow	Github	jsyoon0823/GAIN
RGAN Esteban et al.	Python, Tensorflow	Github	ratschlab/RGAN
GluGAN Zhu et al.	Jupyter Notebook, Python, Tensirflow	BitBucket	deep-learning-healthcare/

Table 10: Dataset used in the publications

Dataset	Link
SPRINT Clinical Trial Data (Wright Jr et al., 2016)	SPRINT Data Analysis Challenge
Coalition Against Major diseases Online data Repository for AD (Neville et al., 2015)	Critical Path Institute (C-Path)
Philips eICU (Pollard et al., 2018)	Physionet (Goldberger et al., 2000)
Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) (Johnson et al., 2016)	MIMIC Physionet (Goldberger et al., 2000)
Vanderbilt University Medical Center Synthetic Derivative (Roden et al., 2008)	BioVU
UC Irvine Machine Learning Repository (Dua and Graff, 2019)	UCI ML repository
Ward2ICU (Severo et al., 2019)	ArXiv
SEER Cancer Statistics Review (CSR) (Noone et al., 2018)	SEEr Incidence database
PREAGRANT (Fasching et al., 2015)	Seemingly not publicly available. Correspondence address: peter.fasching@uk-erlangen.de
New Zealand National Minimum Dataset (hospital events) (eve)	Data request form
Sutter Palo Alto Medical Foundation (PAMF) (Choi et al., 2017a)	
Heart failure study dataset from Sutter (Choi et al., 2017a)	

8 Conclusion

OHD-GAN Acronyms

1D-CAE 1-dimensional Convolutional AE. 9

AC-GAN Auxiliary Classifier GAN. 6

ADS-GAN Anonymization through data synthesis using GAN. 7

AnomiGAN GANs for anonymizing private medical data. 7, 20

CGAIN Categorical GAIN. 6

CONAN *Co*plementary patter**n** **A**augmentatio**n**. 7

corGAN corGAN. 6, 9, 20

CTGAN Conditional Tabular GAN. 6, 9

cWGAN-GP Conditional WGAN-GP. 6

CWR-GAN Cycle Wasserstein Regression GAN. 6, 10

DP-auto-GAN DP-auto-GAN. 7

ehrGAN Electronic Health Record GAN. 6, 9

EMR-WGAN EMR Wassertein GAN. 7, 9

GAIN Generative Adversarial Imputation Network. 6, 10, 20, 22

GANITE Generative Adversarial Nets for inference of Individualized Treatment Effects. 6, 10

GcGAN CorrNN and T-wGAN. 6

GluGAN Blood Glucose GAN. 7, 20

HealthGAN . 6, 7, 9, 10, 14, 15

HGAN Heterogeneous GAN. 7, 9, 16, 17, 19

MC-ARAE Multi-categorical ARAE. 6

MC-GumbelGAN Multi-categorical Gumbel-softmax GAN. 6

MC-medGAN Multi-categorical medGAN. 6–8, 15

MC-WGAN-GP Multi-categorical WGAN with Gradient Penalty. 6, 8

MedBGAN Boundary-seeking medGAN. 6, 8, 9

medGAN medGAN. 5–9, 15, 16, 22

MedWGAN Wassertein medGAN. 6, 8

PATE-GAN Private Aggregation of Teacher Ensembles (PATE) framework applied to GANs. 6

RadialGAN RadialGAN. 6, 10

RC-GAN Recurrent Convolutional GAN. 5, 9, 10

RGAN Recurrent GAN. 5, 9, 10, 20

RMB Restricted Boltzmann Machine. 10

RSDGM Realistic Synthetic Dataset Generation Method. 7

SC-GAN Sequentially Coupled GAN. 6, 10

SSL-GAN Semi-supervised Learning with a learned ehrGAN. 6

T-wGAN Wassertein T-GAN. 6, 22

WGAN-DP WGAN with DP. 7

WGAN-GP WGAN with Gradient Penalty. 6, 7, 9, 15, 22

WGANTPP WGAN for Temporal Point-processes. 6

Glossary

β-VAE β variational auto-encoder. 12

AD Attribute Disclosure. 5, 7, 14, 15

ADTEP Adversarial Deep Treatment Effect Prediction. 6, 10

AE Autoencoder. 5–10, 22

ARAE Adversarially regularized autoencoder. 6, 22

ARM Association Rule Mining. 14

BGAN Method for training GANs with discrete data that uses the estimated difference measure from the discriminator to compute importance weights for generated samples, enabling back-propagation. Tends to push the generated samples to lie on the decision boundary of the discriminator, which also improves stability of training on continuous data (Hjelm et al., 2017). 6

BN batch-normalization. 5, 7, 9

CAE Convolutional AE. 6

CGAN Conditional GAN. 5–7

CNN Convolutional NN. 6, 7, 9

CorrNN Learns a common representation of two views, taking into account their correlation. See (**??**) text. 6, 22

CQM Clinical Quality Measure. 19

CRMB Conditional Restricted Boltzmann Machine. 5–7

CT The self-training and co-training methods use classifiers first trained on the portion of labelled data to predict the labels of unlabelled instances. The newly labelled samples with the highest confidence are added to the labelled set to retrain the classifiers. The process is repeated iteratively. In the words of (Yu et al., 2019), *"[...] co-training splits the features of labeled set into two sub-sets as two views, which are conditionally independent. Two classifiers are trained on two sub-sets respectively, and classify the unlabeled set with pseudo labeled. Then, the most confident unlabeled data*

determined by one classifier is fed into another classifier as additional pseudo labeled data for further training." (Yu et al., 2019). text=CT, first=Co-training (CT). 6, 9

Cycle-GAN Cycle-consistent GAN. 6, 10, 20

digital bio-markers As opposed to classical bio-markers, which are broadly defined as any chemical, physical or biological indication of a patient's state the can be measured and are reproducible (**?**). Digital bio-markers are a trend emerging from the ubiquity of personal electronics devices which are said to have great potential as equivalent indicators, described as "[...] objective, quantifiable, physiological, and behavioural measures that are collected by sensors embedded in portable, wearable, implantable, or digestible devices." (**?**) text=digital bio-markers, first=digital bio-markers. 5

DLE Drug Laboratory Effects refer the changes that a patient's medication can induce on medical laboratory analyses such as diagnostic tests, leading to misinterpretations and errors (Van Balveren et al., 2018). Merely keeping track of the large quantity of known interactions is still problematic and the number of possible combination is immense. Moreover the effects vary according to each patient physiology. Yahi et al. made use of GANs to predict these effect on an personalized basis (Yahi et al., 2017). See also ITE. text. 5

DO Privacy preservation method. See (Yale et al., 2019a) based on (**??**). text. 7

DP Differential privacy. 2, 4, 6, 7, 14, 22, 23

DP-SGD Differential private stochastic gradient descent. 5, 7

DSM Domain Specific Measure. 12

DT The discriminator is tested on batches of synthetic data produced by other methods to asses the possibility of over-fitting, see (Yale et al., 2019a). text. 6

DWP Dimension-wise prediction. 14

EHR Electronic Health Record. 1, 6, 7, 10, 12, 18, 20

FFN Feed-forward Network. 9, 14

FullBB A MI attack setting where an attacker has now knowledge of the internal workings of the generator, but can only sample from it. text. 7

GAN Generative Adversarial Network. 1–12, 15–17, 19, 20, 22, 23

GRU Gated Recurrent Unit. 7

Gumbel-GAN Gumbel-Softmax GAN. 6

HI Health Informatics. 1, 2

ICU Intensive Care Unit. 10

IoT Internet of Things. 5, 6

ITE Given a patient and what we know about the person's medical history and state, and the probability of various possible outcomes of disease progression. The aim of Individualized Treatment effects is to estimate the consequences of administering a particular treatment and their likely-hood. The task is made particularly complex due the fact that for any given individual, the decision can only

be made once. In other words, paired samples are lacking, making it impossible to compare the outcomes directly (**?**). text. 5, 6, 10

KLD Kullback-Leibler divergence. 12

KS Kolmogorov-Smirnov. 12

LN Layer normalisation. 7, 9

LR Logistic-regression. 10, 12

LSTM Long Short-term Memory. 5, 6, 9, 10

MaR Given a dataset with missing entries , the missingness depends only on the observed variables (Yoon et al., 2018b). text. 4, 8, 10

MB-Avg Alternative to mini-batch discrimination that performs well on categorical features to cope with mode collapse, see (**?**) text. 5

MCaR Missing Completely at Random, description=Given a dataset with missing entries, the missingness is not dependant on any of the variables, thus occurs completely at random (Yoon et al., 2018b). text. 6

MI See PD. text. 6, 7, 14

ML Machine Learning. 1, 16, 17

MMD Maximum Mean Discrepency. 12

MSN Per feature, a variational Gaussian mixture model is used to estimate the number of modes and fit a Gaussian mixture. A one-hot vector indicating the mode, and a scalar indicating the value within the mode is produced. See (Xu et al., 2019).. 6

NN Neural Network. 2, 5, 6

NN-AA "Compares the distance from one point in a target distribution T, to the nearest point in a source distribution S, to the distance to the next nearest point in the target distribution." See (Yale et al., 2019a). text. 6, 7

OHD Observational Health Data. 1–4, 7, 16–18

OHD-GAN GANs for Observation Health Data. 3

PartBB Similar to to the FullBB setting with the attacker having the additional knowledge about the latent input z. text. 7

PATE Differential privacy method, best described by **?**: "The approach combines, in a black-box fashion, multiple models trained with disjoint datasets, such as records from different subsets of users. Because they rely directly on sensitive data, these models are not published, but instead used as "teachers" for a "student" model. The student learns to predict an output chosen by noisy voting among all of the teachers, and cannot directly access an individual teacher or the underlying data or parameters. The student's privacy properties can be understood both intuitively (since no single teacher and thus no single dataset dictates the student's training) and formally, in terms of differential privacy." (**??**) text. 6

PCA Principal Component Analysis. 13, 16

PD Broadly, a Membership Inference attack aims to determine if a particular record was used to train a machine learning model (Chen et al., 2019b). There is no canonical process by which an attack is conducted, nor specification of the data assets initially in possession of the attacker. Attacks range from completely FullBB where the attacker can only query data from the model, to WBA where the model and its parameters all fully exposed. For a comprehensive taxonomy of MIA, refer to Chen et al.; **?**. text. 5, 7

PD Probability Distribution. 8, 12

PL Difference of NN-AA on the test set and on the training set. See (Yale et al., 2019a). text. 6, 7

PM According to the NIH, Precision Medicine or "Personalized medicine is an emerging practice of medicine that uses an individual's genetic profile to guide decisions made in regard to the prevention, diagnosis, and treatment of disease. Knowledge of a patient's genetic profile can help doctors select the proper medication or therapy and administer it using the proper dose or regimen." (**?**) text. 5

RDP Renyi Differential Privacy. 7, 12

reidentification attack See MI text. 4

RF Random Forest. 12

RMSE Root Mean-Squared Rrror. 14

RNN Recurrent NN. 5, 7

RR Reproduction rate. 14

SC shortcut connections. 5

SD Synthetic Data. 1–4, 6, 7, 11, 12, 14, 15, 17

SDV Generative model for relational database based on Gaussian Copulas (Patki et al., 2016). One of the few publications treating multi-relation tables in their original form (to out knowledge the only), and has attracted a fair readership. See Github sdv-dev/SDV.. 7

SSL Semi-supervised learning refers to a type of ML algorithm training procedure. Where in supervised learning all the data points are labelled and the algorithm is trained conditionally, and in unsupervised learning the data is unlabelled leaving the algorithm to discover patterns in the data, in semi-supervised learning only a small portion of the data is labelled. There is multitude of glsssl algorithms and application, which generally involve learning from the labelled data points to gather information from the unlabelled. (**?**) text. 4

SSL Semi-supervised learning. 6, 9, 10

ST The self-training and co-training methods use classifiers first trained on the portion of labelled data to predict the labels of unlabelled instances. The newly labelled samples with the highest confidence are added to the labelled set to retrain the classifiers. The process is repeated iteratively. In the words of Yu et al., *"[...] a classifier is initially trained on the small labeled set, and the trained classifier is used to classify the unlabeled set, which is assigned with pseudo labels. After that, the part of unlabeled set with the most confident pseudo labels are selected, and added into the labeled set. The classifier iteratively trains itself with the labeled data and selected unlabeled data."* (Yu et al., 2019). text. 6, 9

SVM Support Vector Machine. 12

T-GAN Training technique to stabilise training. Allows the introduction of real sample information into the process of training the the generator. See (**??**) text. 6, 23

t-SNE The t-Distributed Stochastic Neighbor Embedding clustering algorithm is a nonlinear dimensionality reduction technique commonly applied to high-dimensional data. See **?**. text. 5, 11

TbS To deal with the imbalance of values in categorical featues, during training the data is resampled in a way that all the categories from discrete attributes are sampled evenly, without inducing bias and so as to recover real data distribution. See (Xu et al., 2019) for a step-by-step spefication. text. 6

TRTS Train on synthetic, test on real. 14

TSTR Train on real, test on synthetic. 14

VCD Variational contrastive divergence. 6, 9

WBA Similar to the PartBB and FullBB settings, but with the attacker having full knowledge of the generator internals, including gradient information.. 7

WGAN Wassertein GAN. 6–8, 22, 23

References

Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, Alexander Michel, Seleen Ong, Jill P Pell, Mary Ross Southworth, Wendy Gattis Stough, Martin Thoenes, Faiez Zannad, and Andrew Zalewski. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9, aug 2016. doi: 10.1007/s00392-016-1025-6. URL https://doi.org/10.1007{%}2Fs00392-016-1025-6.

OHDSI. *The Book of OHDSI*. 12 # feb 2020. URL https://ohdsi.github.io/TheBookOfOhdsi/.

Robert S. Rudin, Mark W. Friedberg, Paul Shekelle, Neel Shah, and David W. Bates. Getting Value From Electronic Health Records: Research Needed to Improve Practice. *Annals of Internal Medicine*, 172(11_Supplement):S130–S136, jun 2020. doi: 10.7326/m19-0878. URL https://doi.org/10.7326%2Fm19-0878.

E Capobianco. Imprecise Data and Their Impact on Translational Research in Medicine. *Front Med (Lausanne)*, 7:82, 2020.

Jonathan Ullman and Salil Vadhan. PCPs and the hardness of generating private synthetic data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6597 LNCS, pages 400–416, 2011. ISBN 9783642195709. doi: 10.1007/978-3-642-19571-6_24.

David Enthoven and Zaid Al-Ars. An overview of federated deep learning privacy attacks and defensive strategies, 2020.

Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim. Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review. jul 2020. URL http://arxiv.org/abs/2007.10760.

Byeong Soo Kim, Bong Gu Kang, Seon Han Choi, and Tag Gon Kim. Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system. *SIMULATION*, 93(7):579–594, jun 2017. doi: 10.1177/0037549717692866. URL https://doi.org/10.1177%2F0037549717692866.

David Hand. What is the Purpose of Statistical Modelling? *Harvard Data Science Review*, 1(1), jun 2019. doi: 10.1162/99608f92.4a85af74. URL https://hdsr.mitpress.mit.edu/pub/9qsbf3hz.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, aug 2017. doi: 10.1093/jamia/ocx079. URL https://doi.org/10.1093%2Fjamia%2Focx079.

Junqiao Chen, David Chun, Milesh Patel, Epson Chiang, and Jesse James. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Medical Informatics and Decision Making*, 19(1), mar 2019a. doi: 10.1186/s12911-019-0793-0. URL https://doi.org/10.1186%2Fs12911-019-0793-0.

D Rankin, M Black, R Bond, J Wallace, M Mulvenna, and G Epelde. Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. *JMIR Med Inform*, 8:e18910, Jul 2020.

Ian Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes,

N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

X Yi, E Walia, and P Babyn. Generative adversarial network in medical imaging: A review. *Med Image Anal*, 58:101552, Dec 2019a.

Tonghe Wang, Yang Lei, Yabo Fu, Walter J. Curran, Tian Liu, and Xiaofeng Yang. Medical Imaging Synthesis using Deep Learning and its Clinical Applications: A Review. apr 2020a. URL http://arxiv.org/abs/2004.10322.

S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises. aug 2020. URL http://arxiv.org/abs/2008.09104.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, jun 2018a. doi: 10.1093/jamia/ocy068. URL https://doi.org/10.1093%2Fjamia%2Focy068.

Cristobal Esteban, Stephanie L Hyland, and Gunnar Ratsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, nov 2017. doi: 10.1109/icdm.2017.93. URL https://doi.org/10.1109%2Ficdm.2017.93.

Edward Choi, Siddharth Biswal, Bradley Malin, Duke Jon, Walter F Stewart, and Jimeng Sun. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR, 2017a.

Alexandre Yahi, Rami Vanguri, Noemie Elhadad, and Nicholas P Tatonetti. Generative adversarial networks for electronic health records: a framework for exploring and evaluating methods for predicting drug-induced laboratory test trajectories. *arXiv preprint arXiv:1712.00164*, 2017.

Google. Google Scholar. URL https://scholar.google.com/.

Clarivate. Trusted publisher-independent citation database - Web of Science Group. URL https://clarivate.com/webofsciencegroup/solutions/web-of-science/.

Prophy. Prophy. URL https://www.prophy.science/.

Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, dec 2019b. doi: 10.1016/j.media.2019.101552. URL https://doi.org/10.1016%2Fj.media.2019.101552.

N Nakata. Recent technical development of artificial intelligence for diagnostic medical imaging. *Jpn J Radiol*, 37:103–108, Feb 2019.

Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. Medical Image Analysis using Convolutional Neural Networks: A Review. *Journal of Medical Systems*, 42(11), oct 2018. doi: 10.1007/s10916-018-1088-1. URL https://doi.org/10.1007%2Fs10916-018-1088-1.

Tirthajyoti Sarkar. Synthetic data generation — a must-have skill for new data scientists, 2018. URL https://towardsdatascience.com/synthetic-data-generation-a-must-have-skill-for-new-data-scientists-915896c0c1ae.

Lu Wang, Wei Zhang, and Xiaofeng He. Continuous Patient-Centric Sequence Generation via Sequentially Coupled Adversarial Learning. In *Database Systems for Advanced Applications*, pages 36–52. Springer International Publishing, 2019. doi: 10.1007/978-3-030-18579-4_3. URL https://doi.org/10.1007%2F978-3-030-18579-4_3.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. feb 2018a.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing Data Imputation using Generative Adversarial Nets, 2018b.

Yinchong Yang, Zhiliang Wu, Volker Tresp, and Peter A. Fasching. Categorical EHR Imputation with Generative Adversarial Nets. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, jun 2019a. doi: 10.1109/ichi.2019.8904717. URL https://doi.org/10.1109%2Fichi.2019.8904717.

Limeng Cui, Siddharth Biswal, Lucas M Glass, Greg Lever, Jimeng Sun, and Cao Xiao. CONAN: Complementary Pattern Augmentation for Rare Disease Detection. *arXiv preprint arXiv:1911.13232*, 2019.

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Matthew McDermott, Tom Yan, Tristan Naumann, Nathan Hunt, Harini S Suresh, Peter Szolovits, and Marzyeh Ghassemi. Semi-supervised biomedical translation with cycle wasserstein regression gans. 2018.

Jinsung Yoon, James Jordon, and Mihaela Schaar. RadialGAN: Leveraging multiple datasets to improve target-specific predictive models using Generative Adversarial Networks. In *International Conference on Machine Learning*, pages 5685–5693. proceedings.mlr.press, 2018c.

Brett K Beaulieu-Jones, Zhiwei Steven Wu, Williams Chris, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *Circ. Cardiovasc. Qual. Outcomes*, 12(7):e005122, jul 2019.

M. K. Baowaly, C. Liu, and K. Chen. Realistic data synthesis using enhanced generative adversarial networks. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 289–292, 2019.

Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 12 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy142. URL https://doi.org/10.1093/jamia/ocy142.

CK Fisher, AM Smith, and JR Walsh. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. *Sci Rep*, 9:13622, Sep 2019.

Daniel Severo, Flávio Amaro, Estevam R Hruschka Jr, and André Soares de Moura Costa. Ward2icu: A vital signs dataset of inpatients from the general ward. *arXiv preprint arXiv:1910.00752*, 2019.

Jonathan R Walsh, Aaron M Smith, Yannick Pouliot, David Li-Bland, Anton Loukianov, and Charles K Fisher. Generating Digital Twins with Multiple Sclerosis Using Probabilistic Neural Networks. *arXiv preprint arXiv:2002.02779*, 2020.

Kieran Chin-Cheong, Thomas Sutter, and Julia E Vogt. Generation of Heterogeneous Synthetic Electronic Health Records using GANs. In Institute for Machine Learning ETH Zurich, editor, *Workshop on Machine Learning for Health (ML4H) at the 33rd Conference on Neural Information Processing*

Systems (NeurIPS 2019), dec 2019. doi: 10.3929/ethz-b-000392473. URL https://doi.org/10.3929/ethz-b-000392473.

Jiebin Chu, Wei Dong, and Zhengxing Huang. Treatment Effect Prediction with Generative Adversarial Networks Using Electronic Health Records. In *SEPDA@ ISWC*, pages 53–57, 2019.

J. Wolterink, Anna M. Dinkla, M. H. Savenije, P. Seevinck, C. V. D. Berg, and I. Igum. Deep MR to CT Synthesis Using Unpaired Data. In *SASHIMI@MICCAI*, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017a.

Cecilio Angulo Bahon, Juan Antonio Ortega Ramirez, and Luis Gonzalez Abril. Towards a healthcare digital twin. In *Artificial Intelligence Research and Development vol. 319*, pages 312–315. IOS Press, 2019.

Cecilio Angulo, Luis Gonzalez-Abril, Cristobal Raya, and Juan Antonio Ortega. A Proposal to Evolving Towards Digital Twins in Healthcare. In *Bioinformatics and Biomedical Engineering*, pages 418–426. Springer International Publishing, 2020. doi: 10.1007/978-3-030-45385-5_37. URL https://doi.org/10.1007%2F978-3-030-45385-5_37.

Andrea Coravos, Sean Khozin, and Kenneth D Mandl. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ digital medicine*, 2(1):1–5, 2019.

Christopher W Snyder, E Ray Dorsey, and Ashish Atreja. The best digital biomarkers papers of 2017. *Digital Biomarkers*, 2(2):64–73, 2018.

Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Yan Junchi, Le Song, and Hongyuan Zha. Wasserstein Learning of Deep Generative Point Process Models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS#39;17, pages 3250–3259, Red Hook, NY, USA, 2017. Curran Associates Inc.

Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, Andrew Yale, and Kristin P Bennett. Synthetic Event Time Series Health Data Generation. Technical report.

Ramiro Camino, Christian Hammerschmidt, and Radu State. Generating Multi-Categorical Samples with Generative Adversarial Networks. jul 2018.

Tensorflow community. tensorflow/privacy: Library for training machine learning models with privacy for training data, 2020. URL https://github.com/tensorflow/privacy.

James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. PATE-GaN: Generating synthetic data with differential privacy guarantees. In *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR, 2019.

Amirsina Torfi and Mohammadreza Beyki. Generating Synthetic Healthcare Records Using Convolutional Generative Adversarial Networks. 2019.

Piper Jackson and Marco Lussetti. Extending a Generative Adversarial Network to Produce Medical Records with Demographic Characteristics and Health System Use. In *2019 IEEE 10th Annual Information Technology Electronics and Mobile Communication Conference (IEMCON)*. IEEE, oct 2019. doi: 10.1109/iemcon.2019.8936168. URL https://doi.org/10.1109%2Fiemcon.2019.8936168.

Kezi Yu, Yunlong Wang, Yong Cai, Cao Xiao, Emily Zhao, Lucas Glass, and Jimeng Sun. Rare Disease Detection by Sequence Modeling with Generative Adversarial Networks. 2019.

Yun Yang, Fengtao Nan, Po Yang, Qiang Meng, Yingfu Xie, Dehai Zhang, and Khan Muhammad. GAN-Based semi-supervised learning approach for clinical decision support in health-IoT platform. *IEEE Access*, 7:8048–8057, 2019b. ISSN 21693536. doi: 10.1109/ACCESS.2018.2888816.

Fan Yang, Zhongping Yu, Yunfan Liang, Xiaolu Gan, Kaibiao Lin, Quan Zou, and Yifeng Zeng. Grouped Correlational Generative Adversarial Networks for Discrete Electronic Health Records. In *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, pages 906–913. IEEE, nov 2019c. ISBN 9781728118673. doi: 10.1109/BIBM47256.2019.8983215. URL https://doi.org/10.1109{%}2Fbibm47256.2019.8983215.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling Tabular data using Conditional GAN. In H Wallach, H Larochelle, A Beygelzimer, F d'Alche Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7335–7345. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8953-modeling-tabular-data-using-conditional-gan.pdf.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Privacy preserving synthetic health data. In *ESANN 2019 - Proceedings, 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 465–470, Bruges, Belgium, apr 2019a. ISBN 9782875870650. URL https://hal.inria.fr/hal-02160496.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, apr 2020. doi: 10.1016/j.neucom.2019.12.136. URL https://doi.org/10.1016%2Fj.neucom.2019.12.136.

Uthaipon Tantipongpipat, Chris Waites, Digvijay Boob, Amaresh Ankit Siva, and Rachel Cummings. Differentially Private Mixed-Type Data Generation For Unsupervised Learning. *undefined*, 2019. URL http://arxiv.org/abs/1912.03250.

Ho Bae, Dahuin Jung, Hyun-Soo Choi, and Sungroh Yoon. Anomigan: Generative adversarial networks for anonymizing private medical data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 25:563–574, 2020a. ISSN 2335-6928. URL http://europepmc.org/abstract/MED/31797628.

Taiyu Zhu, Xi Yao, Kezhi Li, Pau Herrero, and Pantelis Georgiou. Blood Glucose Prediction for Type 1 Diabetes Using Generative Adversarial Networks. In Kerstin Bach, Razvan C Bunescu, Cindy Marling, and Nirmalie Wiratunga, editors, *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020*, volume 2675 of *{CEUR} Workshop Proceedings*, pages 90–94. CEUR-WS.org, 2020. URL http://ceur-ws.org/Vol-2675/paper15.pdf.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models, 2019b.

Kieran Chin-Cheong, Thomas Sutter, and Julia E. Vogt. Generation of Differentially Private Heterogeneous Electronic Health Records, 2020.

Ziqi Zhang, Chao Yan, Diego A. Mesa, Jimeng Sun, and Bradley A. Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *Journal of the American Medical Informatics Association*, 27(1):99–108, jan 2020. ISSN 1527974X. doi: 10.1093/jamia/ocz161. URL https://academic.oup.com/jamia/article/27/1/99/5583723.

Chao Yan, Ziqi Zhang, Steve Nyemba, and Bradley A. Malin. Generating Electronic Health Records with Multiple Data Types and Constraints, 2020.

Eda Bilici Ozyigit, Theodoros N Arvanitis, and George Despotou. Generation of Realistic Synthetic Val-

idation Healthcare Datasets sing Generative Adversarial Networks. *Studies in health technology and informatics*, 272:322–325, 2020.

J Yoon, LN Drumright, and der Schaar M van. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE J Biomed Health Inform*, 24:2378–2388, Aug 2020.

A Goncalves, P Ray, B Soper, J Stevens, L Coyle, and AP Sales. Generation and evaluation of synthetic patient data. *BMC Med Res Methodol*, 20:108, May 2020.

Teodora Sandra Buda, Thomas Cerqueus, John Murphy, and Morten Kristiansen. ReX: Extrapolating relational data in a representative way. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9147, pages 95–107. Springer Verlag, 2015. ISBN 9783319204239. doi: 10.1007/978-3-319-20424-6_10. URL https://link.springer.com/chapter/10.1007/978-3-319-20424-6{_}10.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The Synthetic Data Vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, oct 2016. doi: 10.1109/dsaa.2016.49. URL https://doi.org/10.1109%2Fdsaa.2016.49.

J. W. Zhang and Y. C. Tay. Dscaler: Synthetically scaling a given relational database. In *Proceedings of the VLDB Endowment*, volume 9, pages 1671–1682. Association for Computing Machinery, oct 2015. doi: 10.14778/3007328.3007333. URL https://dl.acm.org/doi/10.14778/3007328.3007333.

Y. C. Tay, Bing Tian Dai, Daniel T. Wang, Eldora Y. Sun, Yong Lin, and Yuting Lin. UpSizeR: Synthetically scaling an empirical relational database. *Information Systems*, 38(8):1168–1183, nov 2013. ISSN 03064379. doi: 10.1016/j.is.2013.07.004.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2016.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, 2017.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

Henning Petzka, Asja Fischer, and Denis Lukovnikov. On the regularization of Wasserstein GaNs. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, sep 2018. URL https://arxiv.org/abs/1709.08894.

R Devon Hjelm, Athul Paul Jacob, Tong Che, Adam Trischler, Kyunghyun Cho, and Yoshua Bengio. Boundary-Seeking Generative Adversarial Networks, 2017.

Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. Generating High-fidelity, Synthetic Time Series Datasets with DoppelGANger. sep 2019. URL http://arxiv.org/abs/1909.13403.

Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Zongben Xu, and Jerry Prince. Unpaired Brain MR-to-CT Synthesis using a Structure-Constrained CycleGAN, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017b. doi: 10.1109/iccv.2017.244. URL https://doi.org/10.1109%2Ficcv.2017.244.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation, 2017b.

Ali Borji. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179: 41–65, feb 2019. ISSN 1090235X. doi: 10.1016/j.cviu.2018.10.009.

Shota Harada, Hideaki Hayashi, and Seiichi Uchida. Biosignal Generation and Latent Variable Analysis with Recurrent Generative Adversarial Networks. *IEEE Access*, 7:144292–144302, may 2019. ISSN 21693536. doi: 10.1109/ACCESS.2019.2934928.

Han Jiawei. CS412 2.4.8 Kullback-Leibler Divergence. Technical report, Univ. of Illinois at Urbana-Champaign, 2018. URL http://hanj.cs.illinois.edu/cs412/.

Jasmijn A. Van Balveren, Wilhelmine P.H.G. Verboeket-Van De Venne, Lale Erdem-Eraslan, Albert J. De Graaf, Annemarieke E. Loot, Ruben E.A. Musson, Wytze P. Oosterhuis, Martin P. Schuijt, Heleen Van Der Sijs, Rolf J. Verheul, Holger K. De Wolf, Ron Kusters, and Rein M.J. Hoedemakers. Impact of interactions between drugs and laboratory test results on diagnostic test interpretation-a systematic review, dec 2018. ISSN 14374331. URL http://orcid.org/0000-0001-5451-3010.

Wikipedia contributors. Jaccard index - Wikipedia. URL https://en.wikipedia.org/wiki/Jaccard{_}index.

Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using GAN for improved liver lesion classification. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, apr 2018. doi: 10.1109/isbi.2018.8363576. URL https://doi.org/10.1109%2Fisbi.2018.8363576.

Ho Bae, Dahuin Jung, Hyun Soo Choi, and Sungroh Yoon. AnomiGAN: Generative adversarial networks for anonymizing private medical data. *Pacific Symposium on Biocomputing*, 25 (2020):563–574, 2020b. ISSN 23356936. doi: 10.1142/9789811215636_0050. URL https://www.ncbi.nlm.nih.gov/pubmed/31797628http://psb.stanford.edu/psb-online/proceedings/psb20/abstracts/2020{_}p563.htmlhttps://doi.org/10.1142/9789811215636{_}0050http://arxiv.org/abs/1901.11313AllPapers/B/Baeetal.2020-AnomiGAN-GenerativeAdve.

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993. ISSN 0163-5808. doi: 10.1145/170036.170072. URL https://doi.org/10.1145/170036.170072.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Privacy Preserving Synthetic Health Data. In *ESANN 2019 - European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, Apr 2019b. URL https://hal.inria.fr/hal-02160496.

S Xiao, H Xu, J Yan, M Farajtabar, X Yang, and others. Learning conditional generative models for temporal point processes. *Thirty-Second AAAI*, 2018b.

Bingzhe Wu, Shiwan Zhao, Chaochao Chen, Haoyang Xu, Li Wang, Xiaolu Zhang, Guangyu Sun, and Jun Zhou. Generalization in Generative Adversarial Networks: A Novel Perspective from Privacy Protection. In H Wallach, H Larochelle, A Beygelzimer, F Alche-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 307–317. Curran Associates, Inc., 2019.

Sumit Mukherjee, Yixi Xu, Anusua Trivedi, and Juan Lavista Ferres. Protecting GANs against privacy attacks by preventing overfitting. dec 2019.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models, 2017.

Steven M Bellovin, Preetam K Dutta, and Nathan Reitinger. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22:1, 2019.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Assessing Privacy and Quality of Synthetic Health Data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*. Association for Computing Machinery, 2019c.

Y. Bengio. Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1):1–127, 2009. doi: 10.1561/2200000006. URL https://doi.org/10.1561%2F2200000006.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504, 2016.

Yulia Rubanova, Tian Qi Chen, and David K Duvenaud. Latent Ordinary Differential Equations for Irregularly-Sampled Time Series. In *Advances in Neural Information Processing Systems*, pages 5321–5331, 2019.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems*, pages 7377–7388, 2019.

J Fukae, M Isobe, T Hattori, Y Fujieda, M Kono, N Abe, A Kitano, A Narita, M Henmi, F Sakamoto, Y Aoki, T Ito, A Mitsuzaki, M Matsuhashi, M Shimizu, K Tanimura, K Sutherland, T Kamishima, T Atsumi, and T Koike. Convolutional neural network for classification of two-dimensional array images generated from clinical information may support diagnosis of rheumatoid arthritis. *Sci Rep*, 10:5648, Mar 2020.

Teus H. Kappen, Jonathan P. Wanderer, Linda M. Peelen, Karel G. M. Moons, and Jesse M. Ehrenfeld. Prediction Model for In-hospital Mortality Should Accurately Predict the Risks of Patients Who Are Truly at Risk. *Anesthesiology*, 125(4):815–816, oct 2016. doi: 10.1097/aln.0000000000001269. URL https://doi.org/10.1097%2Faln.0000000000001269.

Teus H. Kappen and Linda M. Peelen. Prediction models. *Current Opinion in Anaesthesiology*, 29 (6):717–726, dec 2016. doi: 10.1097/aco.0000000000000386. URL https://doi.org/10.1097%2Faco.0000000000000386.

Sicco Verwer, Remi Eyraud, and Colin de la Higuera. PAutomaC: a probabilistic automata and hidden Markov models learning competition. *Machine Learning*, 96(1-2):129–154, oct 2013. doi: 10.1007/s10994-013-5409-9. URL https://doi.org/10.1007%2Fs10994-013-5409-9.

Yang Liu, Eunice Jun, Qisheng Li, and Jeffrey Heer. Latent space cartography: Visual analysis of vector space embeddings. *Computer Graphics Forum*, 38(3):67–78, jun 2019. ISSN 14678659. doi: 10.1111/cgf.13672. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13672.

Gregory P. Way, Michael Zietz, Vincent Rubinetti, Daniel S. Himmelstein, and Casey S. Greene. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biology*, 21(1), may 2020. ISSN 1474760X. doi: 10.1186/s13059-020-02021-3. URL https://pubmed.ncbi.nlm.nih.gov/32393369/.

Lefteris Koumakis. Deep learning models in genomics; are we there yet?, jan 2020. ISSN 20010370. URL /pmc/articles/PMC7327302/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7327302/.

Lujia Chen, Chunhui Cai, Vicky Chen, and Xinghua Lu. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinformatics*, 17(1):S9, jan 2016. ISSN 14712105. doi: 10.1186/s12859-015-0852-1. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0852-1.

Shuo Wang, Shangyu Chen, Tianle Chen, Surya Nepal, Carsten Rudolph, and Marthie Grobler. Generating Semantic Adversarial Examples via Feature Manipulation. jan 2020b. URL http://arxiv.org/abs/2001.02297.

Wenhao Ding, Mengdi Xu, and Ding Zhao. CMTS: A Conditional Multiple Trajectory Synthesizer for

Generating Safety-Critical Driving Scenarios. pages 4314–4321. Institute of Electrical and Electronics Engineers (IEEE), sep 2020. doi: 10.1109/icra40945.2020.9197145.

Ziqiang Li, Rentuo Tao, Hongjing Niu, and Bin Li. Interpreting the Latent Space of GANs via Correlation Analysis for Controllable Concept Manipulation. may 2020. URL http://arxiv.org/abs/2006.10132.

K Thorlund, L Dron, JJH Park, and EJ Mills. Synthetic and External Controls in Clinical Trials - A Primer for Researchers. *Clin Epidemiol*, 12:457–467, 2020.

R Etzioni, DF Penson, JM Legler, Tommaso D di, R Boer, PH Gann, and EJ Feuer. Overdiagnosis due to prostate-specific antigen screening: lessons from U.S. prostate cancer incidence trends. *J Natl Cancer Inst*, 94:981–90, Jul 2002.

Christopher V Cosgriff, Daniel K Ebner, and Leo Anthony Celi. Data sharing in the era of COVID-19. *The Lancet Digital Health*, 2(5):e224, may 2020. doi: 10.1016/s2589-7500(20)30082-0. URL https://doi.org/10.1016%2Fs2589-7500%2820%2930082-0.

Regina Becker, Adrian Thorogood, Johan Ordish, and Michael J.S. Beauvais. COVID-19 Research: Navigating the European General Data Protection Regulation. *SSRN Electronic Journal*, 2020. doi: 10.2139/ssrn.3593579. URL https://doi.org/10.2139%2Fssrn.3593579.

Stuart McLennan, Leo Anthony Celi, and Alena Buyx. COVID-19: Putting the General Data Protection Regulation to the Test (Preprint). apr 2020. doi: 10.2196/preprints.19279. URL https://doi.org/10.2196%2Fpreprints.19279.

M Ghassemi, T Naumann, P Schulam, AL Beam, IY Chen, and R Ranganath. A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Jt Summits Transl Sci Proc*, 2020:191–200, 2020.

Koustuv Sinha, Joelle Pineau, Jessica Forde, Jesse Dodge, and Robert Stojnic. Papers with Code : the latest in machine learning, 2020. URL https://paperswithcode.com/https://paperswithcode.com.

Jackson T Wright Jr, Paul K Whelton, and David M Reboussin. A randomized trial of intensive versus standard blood-pressure control. *The New England journal of medicine*, 374(23):2294, 2016.

Jon Neville, Steve Kopko, Steve Broadbent, Enrique Aviles, Robert Stafford, Christine M. Solinsky, Lisa J. Bain, Martin Cisneroz, Klaus Romero, and Diane Stephenson and. Development of a unified clinical trial database for Alzheimer's disease. *Alzheimer's & Dementia*, 11(10):1212–1221, feb 2015. doi: 10.1016/j.jalz.2014.11.005. URL https://doi.org/10.1016%2Fj.jalz.2014.11.005.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*, 5:180178, 2018.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank PhysioToolkit and PhysioNet. *Circulation*, 101(23), jun 2000. doi: 10.1161/01.cir.101.23.e215. URL https://doi.org/10.1161%2F01.cir.101.23.e215.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III a freely accessible critical care database. *Scientific Data*, 3(1), may 2016. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038%2Fsdata.2016.35.

DM Roden, JM Pulley, MA Basford, GR Bernard, EW Clayton, JR Balser, and DR Masys. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clinical Pharmacology & Therapeutics*, 84(3):362–369, may 2008. doi: 10.1038/clpt.2008.89. URL https://doi.org/10.1038%2Fclpt.2008.89.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2019. URL http://archive.ics.uci.edu/ml.

AM Noone, N Howlader, M Krapcho, D Miller, A Brest, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, et al. Cronin (eds) KA. SEER Cancer Statistics Review. 1975–2015, National Cancer Institute, 2018.

P. Fasching, S. Brucker, T. Fehm, F. Overkamp, W. Janni, M. Wallwiener, P. Hadji, E. Belleville, L. Häberle, F.-A. Taran, D. Lüftner, M. Lux, J. Ettl, V. Müller, H. Tesch, D. Wallwiener, and A. Schneeweiss. Biomarkers in Patients with Metastatic Breast Cancer and the PRAEGNANT Study Network. *Geburtshilfe und Frauenheilkunde*, 75(01):41–50, feb 2015. doi: 10.1055/s-0034-1396215. URL https://doi.org/10.1055%2Fs-0034-1396215.

National Minimum Dataset (hospital events). https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/collections/national-minimum-dataset-hospital-events. URL https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/collections/national-minimum-dataset-hospital-events. Accessed on Sun, September 06, 2020.