Synthetic Observational Health Data with GANs: from slow adoption to a boom in medical research and ultimately digital twins?

Georges-Filteau, Jeremy
Radboud University, The Hyve
jeremy@thehyve.nl

Cirillo, Elisa
The Hyve
elisa@thehyve.nl

December 16, 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 (HI) and medical research. Vast potential is unexploited because of the fiercely private nature of patient-related data and regulations to protect it.

Generative Adversarial Networks (GANs) have recently emerged as a groundbreaking way to learn generative models that produce realistic synthetic data. They have revolutionized practices in multiple domains such as self-driving cars, fraud detection, digital twin simulations in industrial sectors, and medical imaging.

The digital twin concept could readily apply to modelling and quantifying disease progression. In addition, GANs posses many capabilities relevant to common problems in healthcare: lack of data, class imbalance, rare diseases, and preserving privacy. Unlocking open access to privacy-preserving OHD could be transformative for scientific research. In the midst of coronavirus disease 2019 (COVID-19) (Cucinotta and Vanelli, 2020), the healthcare system is facing unprecedented challenges, many of which of are data related for the reasons stated above.

Considering these facts, publications concerning GAN applied to OHD seemed to be severely lacking. To uncover the reasons for this slow adoption, we broadly reviewed the published literature on the subject. Our findings show that the properties of OHD were initially challenging for the existing GAN algorithms (unlike medical imaging, for which state-of-the-art model were directly transferable) and the evaluation synthetic data lacked clear metrics.

We find more publications on the subject than expected, starting slowly in 2017, and since then at an increasing rate. The difficulties of OHD remain, and we discuss issues relating to evaluation, consistency, benchmarking, data modelling, and reproducibility.

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 (Abedtash et al., 2020).

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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; Rankin 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 (Dankar and El Emam, 2012; Cheu et al., 2019; De Cristofaro, 2020). These techniques are fallible and do not prevent reidentification. In fact, 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 we must still explore their feasibility and scalability (Raisaro, 2018). 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; Luo et al., 2020; Lyu et al., 2020).

These conditions restrict access to OHD to professionals with academic credentials and financial resources. 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, or educators lack examples (Laderas et al., 2018).

1.2 Synthetic data

An alternative to traditional privacy-preserving methods is to produce full SD. We categorize methods to produce SD as either theory-driven (theoretical, mechanistic or iconic) or 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 or a statistical model representing the causal relationships of a system (Yousefi et al., 2018; Kansal et al., 2018). The Synthea (Walonoski et al., 2017) synthetic patient generator is one such model, in which state transition models[1] produce patient trajectories. It derives the model parameters from aggregate population-level statistics of disease progression and medical knowledge. Such a knowledge-based model depends on prior knowledge of the system, and how much we can intellect about it (Kim et al., 2017; Bonnéry et al., 2019). 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 or reduced utility (Hand, 2019; Rankin et al., 2020). 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, to summarize or describe it (Hand, 2019). There are many statistical modelling approaches to produce SD, but intrinsic assumptions about the data form the basis. They bound their representational power to correlations intelligible to the modeler, being prone to obscure inaccuracies. SD generated by these models hits

[1] Probabilistic model composed of pre-defined states, transitions, and conditional logic.

a ceiling of utility (Rankin et al., 2020). In the ML field, generative models learn an approximation of the multi-modal distribution, from which we can draw synthetic samples (Goodfellow et al., 2014). Generative Adversarial Network (GAN) (Goodfellow et al., 2014) have recently emerged as a groundbreaking approach to 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 affects of GAN to healthcare and science are considerable (Rankin et al., 2020), 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). Well-known characteristics of OHD could explain the relatively slow progress. Primarily, algorithms developed for images and text in other fields were easily repurposed 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, evaluating the realism of synthetic OHD is intuitively complex, a problem that still burdens GANs. In 2017, a few authors first 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 these examples inspired more research, and if so, to gain a comprehensive understanding of approaches to the problem and the techniques involved.

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 terms and operators found in Table 2 form the search query. We included studies reporting the development, application, performance evaluation and privacy evaluation of GAN algorithms to produce OHD. We define OHD as categorical, real-valued, ordinal or binary event data recorded for patient care. We list a more detailed summary of the included and excluded data types in Table 3. The excluded 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
	Time-stamped 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 found 43 publications describing the development or adaption of OHD-GAN, presented in Table 4. We can generalize the data addressed in each of these publications 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. We briefly bring attention to the lack of multi-relational tabular representations, the primary form of EHR, and further discuss the subject in latter sections.

Most efforts propose adaptations of current algorithms to the characteristics and complexities of OHD. These include multi-modality of marginal distributions or non-Gaussian real-valued features, heterogeneity, a combination of discrete and real-valued 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 provides 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 only 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, GAN can generate SD that conveys more information about the real data distribution. Effectively, the real-valued 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., 2019a; 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 would be to seize the concept from

the point of view of image classification, 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

Most authors see SD as the key to unlocking the unexploited value of OHD hindering machine learning, and 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) or education (Laderas et al., 2018). We can broadly describe preserving privacy as reducing the risk of re-identification attack to an acceptable level. It quantifies this level of risk when releasing data anonymized with DP.

Due to its artificial nature, SD is put forward 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 actual people, the significance of such an occurrence is unclear. It is possible to combine both methods, and GAN training according to DP shows evidence of reducing the loss of utility compared to DP alone.

3.2.3 Enabling precision medicine

The application to precision medicine involves 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). Estimating ITEs is persistently hampered by the lack of paired counterfactual samples (Yoon et al., 2018a; Chu et al., 2019). To solve similar problems in 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. The relations learned by the model, its parameters, contain meaningful information if we can learn to harness it. Data-driven algorithms evolve as our understanding of their behavior improves. We incorporate new concepts in the algorithms leading to further understanding, interactivly 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 obscure logic. Incorporating new understanding in the architec-

ture of algorithms shifts 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 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 use 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, algorithms, and techniques	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).	Binary occurences or counts 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 (RV-TS)
Xiao et al.	WGAN for Temporal Point-processes (PPWGAN)	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 (D-TS), sequences of medical codes.
Dash et al.	HealthGAN (HealthGAN)	Sleep patterns, stratification by covariates.	Binary over multiple visits.
2018			

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

Publication	Algorithm(s)	Focus, algorithms and techniques	Data
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).	Multiple categorical variables.
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 RV-TS, 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.
Yoon et al.	Generative Adversarial Imputation Network (GAIN)	Tabular data imputation. Missing Completly at Random (MCaR), CGAN.	Real-valued, tabular 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.	RV-TS of patient state and medication dosage data.
Baowaly et al.	Boundary-seeking medGAN (MedBGAN)	Improving training process. medGAN, Boundary-seeking GAN.	Binary occurences or counts of medical codes.
Baowaly et al.	MedBGAN, Wassertein medGAN (MedWGAN)	Improving training process. medGAN, BGAN, WGAN.	Binary occurences or counts of medical codes.
Severo et al.	Conditional WGAN-GP (cWGAN-GP)	Generation and public release of dataset. Protecting commercial sensitive information. Class imbalance. cWGAN-GP, CGAN.	Physiological RV-TS.
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, real-valued and categorical.
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.
Torfi and Beyki	corGAN (corGAN)	Convolutional NN (CNN) architecture, capturing feature correlations, evaluating realism, privacy evaluation using Membership Inference (MI). 1D-Convolutional AE (CAE).	Binary occurences or counts 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.	Demographic features and binary occurences or counts of medical codes.
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.

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

Publication	Algorithm(s)	Focus, algorithms and techniques	Data
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 or counts of medical codes.
Yang et al.	Categorical GAIN (CGAIN)	Improve on GAIN for categorical variable using fuzzy encoding of the features.	Categorical (multi-class and multi-label) real-valued.
Camino et al.	GAIN, GAIN+Variable Splitting (VS), Variational AE (VAE), VAE+Iterative Imputation (IT), VAE+Backpropagation IT (BP), VAE+VS, VAE+VS+IT, VAE+VS+BP	Benchmark and improve on generative imputation with GAIN and VAE.	Categorical and real-valued. Mostly not OHD
Beaulieu-Jones et al.	Auxiliary Classifier GAN (AC-GAN)	Evaluating if differentially private GANs that is valid reanalysis while ensuring privacy. DP, CGAN.	Physiological RV-TS.
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 real-valued and categorical.
Yale et al.	HealthGAN	Privacy metrics and over-fitting. MI, Nearest-neighbor Adversarial Accuracy (NN-AA), Privacy loss (PL), Discriminator testing (DT)	Categorical demographics, real-valued 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, 3 months intervals.
2020			
Walsh et al.	Adversarially trained CRMB	Digital twins, disease prediction and risk quantification, missingness. CRMB.	Binary, ordinal, categorical, and continuous, 3 months intervals.
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,	Real-valued and categorical. 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 can be viewed as encrypted. DP CNN	Binary occurences of medical codes.
Cui et al.	**C**oplementary 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).	RV-TS of blood glucose measurements, 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 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.

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

Publication	Algorithm(s)	Focus, algorithms and techniques	Data
Camino et al.	-	Initially a comparison GANs and VAEs, but they choose instead to bring attention to the problem of benchmarking. Analysis of problematic, requirements and suggestions. GAIN, Six component GAN (HexaGAN) (Hwang et al., 2019), Missing data IWAE (MIWAE) (Mattei and Frellsen, 2019), Heterogeneous-Incomplete VAE (HI-VAE)(Nazabal et al., 2020), Multiple Imputation Denoising Autoencoders (MIDA) (Gondara and Wang, 2017)	Real-valued and categorical.
Zhang et al.	EMR Wassertein GAN (EMR-WGAN)	Improving training, evaluation metrics, sparsity. WGAN,BN,Layer normalisation (LN), CGAN.	Binary occurences of 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 real-valued.
Ozyigit et al.	Realistic Synthetic Dataset Generation Method (RSDGM)	Exploring the feasibility of various methods to generate synthetic datasets. WGAN	Real-valued 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.	Real-valued and binary.
Goncalves et al.	MC-medGAN	Comparison of GANs with statistical models to generate synthetic data, evaluation metrics. MI, AD.	Categorical and real-valued.

3.3 Data Types and Feature Engineering

No publications made use of OHD in its initial form, patient records in EHR are composed of many related tables (normalized form). The complexity of a model would explode when maintaining referential integrity and statistics between multiple tables. The hierarchy by which 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 published GAN algorithms made to consume normalized database in their original form. In all publications we considered, feature engineering was used to adapt the data to task requirements, or to promising algorithms that fit the data characteristics. They transform the data 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*	- Time-stamped 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. They keep the decoder portion 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.

Early efforts were to improve the performance of medGAN. Among the first, Camino et al. developed MC-medGAN changing 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 (Briefed in Panel 1) with Gradient Penalty (Gulrajani et al., 2017). 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.

Subsequent authors owing to the propensity of OHD to induce mode collapse widely adopted Wasserstein's distance. 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

Panel 1. Wasserstein's distance

In brief, the Wasserstein distance is a measure between two Probability Distributions (PDs) that has the property of always providing a smooth gradient. As the loss function of the discriminator, this property improves training stability and mitigates mode collapse. To make the equation tractable a 1-Lipschitz constraint must be introduced, creating another problem. In the words of the author:

> "Weight clipping is a clearly terrible way to enforce a Lipschitz constraint. If the clipping parameter is large, then it can take a long time for any weights to reach their limit, [...] If the clipping is small, this can easily lead to vanishing gradients [...] However, we do leave the topic of enforcing Lipschitz constraints in a neural network setting for further investigation, and we actively encourage interested researchers to improve on this method." (Arjovsky et al., 2017)

Sometimes this prevented the network from modelling the optimal function, but Gradient penalty, a less restrictive regularization replaced the clipping. (Petzka et al., 2018).

quality, in particular MedBGAN (Baowaly et al., 2019; Baowaly et al., 2018). 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). HealthGAN, based on WGAN-GP, 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 introducing 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 common-sense inconsistencies, such as gender mismatches in medical codes (Yan et al., 2020; Choi et al., 2017a). HGAN enforces constraints 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., 2019)

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

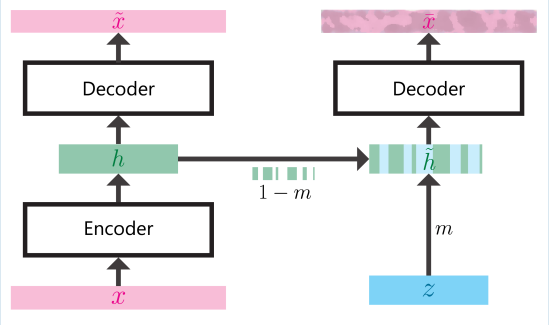
In other efforts, Torfi and Beyki develop corGAN, with a 1-dimensional Convolutional AE (1D-CAE) to capture neighboring feature correlations of the input vectors (Torfi and Beyki, 2019). Chin-Cheong et al. use a Feed-forward Network (FFN) based on Wasserstein's distance to evaluate the capacity of GANs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

3.4.3 Time-series

Esteban et al. devise the LSTM-based RGAN and RC-GAN to generate a regular time-series of physiological measurements from bedside monitors (Esteban et al., 2017). Curiously, the authors dismiss Wasserstein's distance explicitly, and generated each dimension of their time-series independently, where one

Panel 2. Transitional distribution

The ehrGAN generator is trained to decode a random vector z mixed with the latent space representation of a real patient h to produce a synthetic sample $\tilde{x}$ (Che et al., 2017). A standard autoencoder (left) is trained to encode a real patient x to and from a latent representation h, minimizing the reconstruction error with $\bar{x}$. The decoder portion (left) is then trained to produce realistic synthetic samples $\tilde{x}$ from a combination of the random latent vector z and the latent space encoding of a real patient x. The generator thus learns a transition distribution $p(\tilde{x}|x)$ with $x \sim p_{data}(x)$. The amount of contribution of the real sample is controlled by a random mask according to $\tilde{h} = m * z + (1 - m) \cdot h$. This method inspired from Variational Contrastive Divergeance prevents mode collapse by design and learns an information rich transition distribution $p(\tilde{x}|x)$ around real samples x.



would assume they are correlated. They observe a considerable loss of accuracy on their utility metric.

3.5 Task oriented GAN development

3.5.1 Semi-supervised learning

ehrGAN is developed for sequences of medical codes Che et al.. It learns a transitional distribution, combining 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 (See Panel 2). 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 used to augment the minority class in imbalanced datasets, with techniques such as ST and CT. Yang et al. improves on both 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.

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). One encoder-decoder pair per data endpoint 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. The problem is that counterfactual outcomes are never observed or treatment selection is highly biased (Yoon et al., 2018a; McDermott et al., 2018; Walsh et al.,

2020). In GANITE Yoon et al. propose a solution by using 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.

McDermott et al. developed CWR-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). CWR-GAN is a joint regression-adversarial SSL approach inspired by Cycle-GAN. 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. address the problem of data scarcity by designing ADTEP. The algorithm can harness 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 discriminator is tasked with 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).

Like 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., 2019a). 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 can mitigate 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 real-valued features MaR in tabular datasets (Yoon et al., 2018b). In GAIN, the discriminator must classify 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 by Yang et al. to also handle categorical features using fuzzy binary encoding, the same technique employed in HealthGAN. In parallel, Camino et al. apply the same VS technique they used fir medGAN to adapt GAIN and run a benchmark against different types of VAE.

The distribution estimated by a generator 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. As an example, we mentioned generating unobserved counterfactual outcomes (Yoon et al., 2018b), and 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 assess 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 becomes 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 (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., 2020a; Yale et al., 2019a; Yang et al., 2019c; Beaulieu-Jones et al., 2019; Tantipongpipat et al., 2019; Dash et al., 2019).
- **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 quality. 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: comparing the distributions of real and synthetic data as a whole, assessing the marginal and conditional distributions of

features, and 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 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, presented in Table 8, often provide a more convincing indicator of data realism. On the other hand, they mostly lack the interpretability of statistical metrics. We took the liberty of placing these into one of two categories: tasks mostly defined only 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 comparison.
- **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. (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 impaired version. 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 0xCE-VAE, comparing the dimensional mean of the variance or the distance between mode peaks (Zhang et al., 2020). See Section 5.5 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., 2019a; 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., 2020a; Yoon et al., 2020; Walsh et al., 2020; Yale et al., 2019a; Ozyigit et al., 2020; Dash et al., 2019; 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., 2020a; Baowaly et al., 2019; Wang et al., 2019a; 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., 2020a).

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 conducted a privacy risk assessment to evaluate the risk of reidentification. The empirical analyses were based on 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 was also used (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. 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.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.

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 Traditional privacy

Numerous attempts at applying traditional privacy guarantees, such as deferentially-private stochastic gradient descent can also be found in other fields, as well as in healthcare (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.

3.8.3 Moving forward safely

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; Zhu et al., 2020b). Letting go of the negative connotation, we can explore the benefits such as improving generalization, stabilizing learning and building fairer models (Zhu et al., 2020b) and the use of GANs to optimize the trade-off (Chen et al., 2019c).

- 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 pre-trained predictor (Bae et al., 2020a).
- In privGAN (Mukherjee et al., 2019), an adversary is introduced, forcing the generator to produce samples that minimize the risk of MI attacks, 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.

3.8.4 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. This echoes the concept of t-closeness (Li et al., 2010). 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 raises the risk of reproducing the real samples. 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 a tune-able threshold of identifiability. 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 GANs. 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 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 versus 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.

The ongoing Covid-19 pandemic has brought unprecedented levels of cooperation between scientists from around the world. The urgency of obtaining data has highlighted on difficult terms the need for novel ways of sharing and generating data (Bandara et al., 2020; Cosgriff et al., 2020). Global concerted efforts were highly successful, but also required adaptation, with some proposing exemptions from the GDPR (McLennan et al., 2020a). Data sharing was limited to aggregate counts, rather than at the patient level, limiting the depth of analyses.

In the beginning of an epidemic, the scarcity of data can be compensated with synthetic data. Building generative statistical methods in such conditions is a difficult task (Latif et al., 2020). While additional data becomes available to fine-tune the model, so do the number of features and the complexity of the model. This was attempted by Synthea (Walonoski et al., 2017) in the early months of the pandemic, with humble results, nonetheless they were used in many online challenges, hackathons, and conferences.

> The authors state that if one takes "[...] Field Marshall Moltke's notion of "no plan survives contact with the enemy" as true and expands the scope to modeling and simulation, then we might say that "no model survives contact with reality." (Walonoski et al., 2020). We would argue that GANs grow stronger in contact with reality.

Generative models refine their representation as more data is provided and could be combined with current methods of forecasting. When the amount of ground truth data is small, semi-supervised learning simulations can improve the performance of predictors (Dahmen and Cook, 2019). Domain translation, as demonstrated in RadialGAN, would be exceptionally useful to combine datasets from disparate localities. In a recent publication, two different data augmentation techniques provided a significant increase in sensitivity and specificity for the detection of COVID-19 infections, one of which producing SD with a GAN (Sedik et al., 2020).

4.2 Challenges posed by OHD

The challenges posed by health data for GANs are obvious, a number of recurrent factors influence the outcome of efforts to develop them. These problems are not limited to generative algorithms, but also ML in general. For generative models, multi-modality caused the most trouble in achieving a stable training procedure. At the outset, preventing mode collapse attracted the most research efforts, in addition to data combinations of categorical and real-valued features. A rapid succession of efforts aimed at improving medGAN by incorporating the latest machine learning techniques showed continued improvements. However, taken as a whole the efforts were haphazard in their methods and metrics. Often yielding unsurprising results, considering the techniques were known to improve performance across a broad range of applications. This is expected in a new field of application, and more concerted efforts to systematically approach the problems should progressively form.

4.2.1 Feature engineering

We observed that majority of methods included in the review made use of heavily transformed representations of patient records. 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 (1) a timestamp, (2) a medical concept and (3) 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. These properties are not typically considered practical for machine learning.

At varying degrees, depending on the transformations, information is being lost or bias is introduced. For example, when data are reduced by aggregation to one-hot encoding, the complex relationships found in medical data are, for the most part eliminated. Similarly, information is lost when forcing real-valued 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, or the medical codes are truncated to their parent generalizations (Zhang et al., 2020; Choi et al., 2017a).

4.3 From innovation to adoption: Evaluation metrics and benchmarking

Interesting innovations were demonstrated, and progress has good momentum. Their application and adoption will undoubtedly be more sluggish, as has been the case with predictive ML. For good reason, the bar is set high in demonstrating consistent outcomes and ensuring patient safety. While the problem

Panel 3. Representation and visualisation

Ledesma et al. describe the problem of medical data representation and visualization learnedly, from information quality and usefulness, timescales and perception, to user satisfaction and aesthetics. The evaluation of their solution is extensive, detailed and rigorous, done according to the well known Nielsen's heuristics for Human-Computer Interaction (Nielse). Interested readers can find the remainder here Ten Usability Heuristics. While this may seem like total digression towards graphic design, it is rather to illustrate the complexity of aspects to be considered before representing data in a evaluation task.

Principle #2: Match between system and the real world:
The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
—
Principle #4 Consistency and standards
Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

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. Consistent behavior and reproducible results will be required to expect any significant adoption. In regards to evaluation, it is manifest that the choice of optimal metrics and indicators is still being explored. The fact is that the efforts are far from consistent or systematic. As an 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). Overall, none of the evaluation metrics addressed the concept of realism in synthetic data.

4.3.1 Qualitative realism

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., 2019a). It is doubtful that they could in fact be distinguished. Esteban et al. found that participants avoided the median score and were not confident enough to choose either extreme (Esteban et al., 2017).

In their evaluation of medGAN, (Yale et al., 2019b) argue that the positive resemblance of plotted feature distributions is due to the fact that the model's architecture tends to favor reproducing the means and probabilities of each diagnosis column. They note that synthetic data contains samples with an unusually high number of codes, which is not apparent in the plots. Their hypothesis is that these samples are used by the algorithm to discharge the rare medical codes with weak correlation, in an effort to balance the distributions. However, they stated in their experiments that comparing PCA plots of real and synthetic data was nonetheless insightful to get an impression of their behavior (Yale et al., 2020) If visual inspection is to be used, it should be done systematically according established frameworks (See Panel 3)

4.3.2 Quantitative fitness

Reproducing aggregate statistical properties is rather unconvincing evidence that a model has learned to reproduce the complexity of patient health trajectories. In some cases the statistical metrics may be contradictory, such as when the ranking of medical frequencies in the data are wrong, but augmentation leads to improved performance (Che et al., 2017). 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 do not represent a solution. 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 (Zhang et al., 2020).

Utility-based metrics do overall provide a more solid evaluation of data quality. However, they 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. While such consideration was found in sparingly in the publications, extensive research available on the subject of medical information representation. The complexity of health data and its variety make it a considerable, but captivating challenge.

4.3.3 Constraints

As described in Section 3.4.2, HGAN 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. However, the constraints were strict and would be hard to scale. The idea of incorporating knowledge-based constraints in the otherwise naive GAN is in fact gaining attention (See Section 5.4

5 Suggestions of requirements for OHD-GAN development

5.1 Models of appropriate scope and equivalent degree of evaluation

Overall, evaluation methods were superficial or uni-dimensional relative to the scope of the task. As previously discussed, finding convincing and robust evaluation metrics for synthetic health data is an open issue. Weak metrics become a prominent issue when the learning task is broad, loosely defined, constructed for the sole purpose of evaluation, or the scope of application 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 develop robust algorithms for a single data type or modality.

- This makes qualitative evaluation by visual inspection from experts possible and meaningful.
- The behaviour of the model can be assessed straightforwardly
- Conditional models are easier to develop.
- The evaluation metrics should not be defined solely for the purpose but from a peer-reviewed healthcare publication.

A baby learns to crawl, walk and then run. We are in the crawling stage when it comes to applying machine learning.

Dave Waters

5.2 Data-driven architecture

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 of the data at a higher level of abstraction. This type of data-oriented architecture has led to impressive performance for CNN and image data.

Health data presents a different, analogous multi-level structure. As an illustration, a predictive algorithm developed in a hierarchical structure was shown to form representations of EHR that capture the sequential order of visits and co-occurrence of codes within a visit. It 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 [2], such as Electroencephalograms (EEGs) and Electrocardiograms (ECGs) found in EHR data, have shown improved accuracy over discrete time-representations (Rubanova et al., 2019; De Brouwer et al., 2019). 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)

The architecture of OHD-GAN should be engineered to match the data, not the other way around. Data with minimal transformations, to the extent possible. In addition to preventing information loss, this ensures models will reflect the real generative process. Such models are more likely to further our understanding about them and the biological drivers. With deeper understanding, novel architecture of higher complexity will be engineered. Furthermore, the learned statistical distribution is inevitably more meaningful and interpretable, facilitating applications in the healthcare domain and supporting the inference of insights.

Torture the data, and it will confess to anything.

Ronald Coase

5.3 Evolving the patients

As we have seen, OHD-GAN are not exclusively used to produce ”fake” patients, but also to be representative of an particular patient. Common examples are translating between patient states, or producing counterfactuals. It would be interesting to see if combining GAN with what is know as evolutionary computing could produced valuable results. We can think of a GAN transforming the patient data to an alternative state, after which the evolutionary algorithms would optimize this new state in a continuous fashion, as new data about the patient becomes available. Immediately after writing this, a quick search confirms the combination can have impressive results, either in optimizing the evolutionary process (He et al., 2020), exploring the latent space (Schrum et al., 2020), or expanding the information received by the discriminator (Mu et al., 2020).

[2] Those interested in GAN for wavelike data will find many examples (Delaney et al., 2019; Golany and Radinsky, 2019; Ye et al., 2019; Wang et al., 2019b; Singh and Pradhan, 2020; Aznan et al., 2019; Hartmann et al., 2018).

> Unexpected combinations of existing algorithms can harness the in strengths of both, or compensate for lacking, producing performance above the capabilities of both. Evolutionary algorithms are only one particular example, but in fact we've seen a few in this review, including the first that incorporated an AE or the techniques borrowed from the SDV. Mix-and-match, select, repeat is the principle behind any ML model, the notion of meme, and broadly human knowledge... and our existence.

> To me, it is very striking to now understand that their work, described in "ImageNet Classification with deep convolutional neural networks", is the combination of very old concepts (a CNN with pooling and convolution layers, variations on the input data) with several new key insight (very efficient GPU implementation, ReLU neurons, dropout), and that this, precisely this, is what modern deep learning is.
>
> *Andrey Kurenkov (Kurenkov, 2020)*

5.4 Forcing, disciplining or guiding

To build statistical models we define rules and relations that they are forced to optimize when learning. On the other hand, GANs are given free range in a space of possibilities and are disciplined for exploring certain areas, but are provided no explanation.

We build enormous models and let them fight back and forth in a mim-max battle that goes on forever, denying them our valuable knowledge. The idea of introducing human knowledge in the otherwise naive training process has gained some attention.

Posterior regularization is usually used to impose constraints on probabilistic models, but GANs lack the necessary Bayesian component. In the student-teacher model, where a larger model is used to train a smaller one, the process is knowledge distillation. Such models are developed for many applications, such as compression, improving accuracy and accelerating training (Abbasi et al., 2019).

In the field of Reinforcement learning (RL), a mathematical correspondence between Posterior Regularization (PS) and RL led to the probabilistic Posterior Regularization (PR) framework Inverse RL (IRL) that seeks to learn a reward function from expert demonstrations. This was followed by approaches capable of learning both the reward function and the policy (Finn et al., 2016; Fu et al., 2018). Hu et al. then demonstrated a correspondence between RLs and GANs. This allowed them to develop a GAN with a constraint-based learning objective (Hu et al., 2018).

The constraints, seen as a reward function, can be learned by the model through an algorithm involving maximum entropy. This means the known constraints can be input directly or partially and left to be learned automatically. The algorithm consistently improved the speed and quality of training, and accuracy on a few tasks. The approach is exemplified on an image translation task where images of people are transformed from one pose (ex. looking forward) to another (ex. head turned left). The constraint is provided by a pre-trained auxiliary classifier that assigns each pixel to a body part, and is adapted jointly with the GAN. The GAN is rewarded for preserving the mapping in the output image. A performance comparison against unconstrained and fixed-constraint models results in similar training loss and evaluation metric. However when evaluated by humans, the novel approach surpasses the other models on 77% of test cases.

The prospect of GANs being able go incorporate auxiliary information or constraints they can automatically learn to optimize is a golden research opportunity. This would empower them with the prior knowledge until now reserved to model in the category of the same name, while keeping their ability to learn in an unsupervised adversarial framework.

"If you had all the world's information directly attached to your brain, or an artificial brain that was smarter than your brain, you'd be better off."

Sergey Brin

5.5 Interpretability

Even though a few authors attempted to understand the behavior of their models, overall the subject was left largely unmentioned. It is imperative that future experimentation and publication give equal importance to the interpretation of their models and establishing means to do so. In the healthcare domain, black box machine learning models find little adoption, and synthetic data is most often met with dismissal to its validity. The task is not impossible, as for any other opaque system, and in fact experimental sciences in general. The simplest approach is to provide input, observe the output, reformulate our hypotheses, and modify the input accordingly. Repeatedly, to convergence. Fortunately, in this case the internals workings are entirely available, tipping the balance between brute-force, and knowledgeable-driven exploration of the system. In addition, we believe "qualitative" evaluation by visual inspection has much greater potential, still to be defined. What better to define interpretation than a medical professional decoding the hidden relations in data visually.

In theory, the latent space is a lower-dimensional representation of basic concepts that should be directly interpretable. However, in practice these concepts are entangled over multiple nodes. In what is a preliminary, but encouraging proof-of-concept, (Liu et al., 2019) explore how they can use perturbations to reveal patterns in a 0xCE-VAE trained to capture brain structure in mice. By generating a collection of images from a dense interpolation of the latent space, they were able to examine the projective field of latent variables onto the pixels. They found zones of high variance that corresponded to biologically relevant areas. Reversing the experiment, they masked areas of the images and found that many latent factors were not activated by all regions of interest and had localized receptive fields. Whereas complex highly connected regions such as the hippocampus activated almost all latent factors. Curiously, the projective and receptive fields may not be aligned. Numerous other 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., 2016a). 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). However, a recent information-theoretic GAN simplified interpretation greatly by forcing the latent nodes to learn disentangled representations. In addition to adversarial loss, Information Maximizing GAN (InfoGAN) also maximizes the mutual information between small numbers of latent nodes. The result is highly interpretable nodes that represent distinct concepts that can be easily influenced, or in some cases interpolate smoothly between features (Chen et al., 2016b).

5.6 Benchmarking, a priority

It became slowly obvious through the secession of experiments that there is a glaring problem of standardization of evaluation. New algorithms and applications are being demonstrated at an increasing rate. On the contrary, standardized benchmarks, procedures to transform the data, and source has remained scarce, one can hardly compare the models objectively or nominate the best performances. Commendably, Camino et al. are the first to bring attention to this issue in a position paper that provides quantitative arguments. Notably the myriad of ways commonly used datasets are reprocessed, metrics that are not comparable, and hyperparater sweep results, for which no transformation code and optimal values are released and the lack of effort towards reproducibility will only reduce credibility of the field. On a positive note, we've compiled a list of the repositories which were made open-source in Table 9 and a list of the common dataset links can be found in Table 9.

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 (Reiner Benaim et al., 2020). 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 (Qayyum et al., 2020). Medical professionals care foremost about the well-being of their patients and will only consider results obtained from synthetic data if they have the assurance that they are valid (Rankin et al., 2020). A remarkable resource for the purpose of benchmarking is the clinical prediction benchmarks defined on the Medical Information Mart for Intensive Care (MIMIC) data by Harutyunyan et al.. The tasks are clearly defined and the source code to process the data and the algorithms is available (Harutyunyan et al., 2019). We suggest comparing the accuracy of the predictive algorithms applied to the original data versus the synthetic data to be evaluated. However, concerted efforts and informal guidelines that can be agreed upon should be on a regular schedule. We fully support the idea or organized challenges and hackahton proposed by (Camino et al., 2020) and suggest a progressive approach to realizing it.

5.6.1 Ultra-open source, collaborative, publishing communities

In a successful and educative experiment on collaborative writing and crowd-sourcing, an article was entirely written in an open-source GitHub repository. Anyone willing to add their knowledge to the publication was welcome to do so, reaching 30+ authors in 20 countries. Every change proposal is requested for inclusion by a Pull Request, for which R2-3 approvals are necessary. Withing minutes, automated deployment procedures (Github since then released Actions, requiring minimal coding), took care of verifying compliance to guidelines, citation management, DOI registration, and compilation of latex or Markdown. Withing minutes a revised document is released, making the publication a contiguously up-to-date source of knowledge, that can be augmented in the web version with interactive code-books and figures.

Issues can be discussed in the appropriate channels, but most importantly the nature of GitHub ensures attribution of work done, down to a single character. The authors also implemented immutable backup on the blockchain. Since then distributed storage and computation blockchains have reached maturity and could store models, training artefacts, and data for competition at a trivial cost. As an alternative, the Weights and Biases (WandB) platform is a fitting environment, worth a look even for individuals. The traditional publishers have long been touting a makeover of the publication system, changes are slow and trivial, whereas den centralized, person to person, systems have been transforming whole sectors faster than ever.

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, are 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, this has been used before in a competition where participants were given synthetic data from finite state transition machines with known probabilities and tasked to build and learn models that would reproduce those of the original, unseen models. The participants worked according to the Perplexity metric, commonly used in NLP, which 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 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., 2020b). An obvious application is generating additional amounts of data in the early stages of the pandemic, potentially creating opportunities earlier. Synthetic data is not only an opportunity to facilitate the exchange of data, but also to 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: Open-source repositories

Author and algorithm	Repository	Format	Data
Baowaly et al. MedBGAN	baowaly/SynthEHR	Tensorflow	✓
Baowaly et al. MedBGAN, MedWGAN	baowaly/SynthEHR	Tensorflow	✓
Severo et al. cWGAN-GP	3778/Ward2ICU	PyTorch	✖
Torfi and Beyki corGAN	astorfi/cor-gan	PyTorch	✓
Jackson and Lussetti medGAN	marcolussetti/extended-medgan	Tensorflow	✓
Beaulieu-Jones et al. AC-GAN	greenelab/SPRINT_gan	Keras	✓
Xu et al. CTGAN	sdv-dev/TGAN	Tensorflow	✓
Yale et al. HealthGAN	yknot/ESANN2019 Codalab 19365	Tensorflow	✓
Yale et al. HealthGAN	TheRensselaerIDEA/synthetic_data	Tensorflow,	✓
Tantipongpipat et al. DP-auto-GAN	DPautoGAN/DPautoGAN	PyTorch	✓
Bae et al. AnomiGAN	hobae/anomigan	Tensorflow, Keras	✓
Zhu et al. GluGAN	deep-learning-healthcare/glugan	Tensorflow	-
Chen et al. medGAN, WGAN-GP, DC-GAN	DingfanChen/GAN-Leaks	PyTorch	✓
Source code was linked in the publication and could not be found.			
Chin-Cheong et al. , Jordon et al. PATE-GAN, Chu et al. ADTEP,Yu et al. SSL-GAN, Yang et al. CGAN, Yang et al. GcGAN, Yang et al. CGAIN, Walsh et al. Adversarial CRMB, Fisher et al. Adversarial CRMB Cui et al. CONAN,Chin-Cheong et al. WGAN-DP,Zhang et al. EMR-WGAN,Yan et al. HGAN,Ozyigit et al. RSDGM, Yoon et al. ADS-GAN, Goncalves et al. MC-medGAN			

Table 10: Relevant 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 Repository for AD (Neville et al., 2015)	CAMD AD/MCI
American Time Use Survey (ATUS) (of Labor Statistics)	ATUS
Philips eICU (Pollard et al., 2018)	Physionet (Goldberger et al., 2000)
Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) (Johnson et al., 2016a)	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)	On request: peter.fasching@uk-erlangen.de
New Zealand National Minimum Dataset (hospital events) (eve)	Data request form
Sutter Palo Alto Medical Foundation (PAMF) Heart failure study (Choi et al., 2017a)	PAMFRI

8 Conclusion

SD has been a subject of interest for quite some time, with officials seeing enough value to launch longitudinal state-wide endeavours such as the Synthetic Data Project (SDP), funded by the United States Department of Education(USDOE) (Bonnéry et al., 2019). They dismiss a series of anonymization techniques, stating the burden on worker and financial resources, and the privacy guarantees that would not

sufficient for governmental agencies. Issues that have only gained weight with the accumulation of big data, and the number of new sources growing consistently. The questions they hoped to answer at the start of the project in 2016 are still not fully answered (evaluation, scientific validity, legal implications). Their 2019 report on the experience is packed with interesting insights. Noting the distrust people tend to have of synthetic data, they were the ones who first proposed the idea conducting experiments on synthetic data, that could then be confirmed on real data by simply sending the analysis to the data holder (with the logistics described extensively and augmented by a flowchart).

The publication ends with a series case reports. The instances where the data could not satisfy requirements are analyzed with the aim of informing similar projects in the future. However the bulk of reports describe cases where was highly applicable. They concluded by predicting that the cost of generating SD will diminish and that the methods to do so will improve. Their hopes for SD include: easier access for researcher to the wealth of data, increased access providing downstream benefits at the state level, the these benefits encourage others to undertake similar projects that would increase generalizability of findings across states, and a preference for open data.

> "[although some argue for] having secured data centers for administrative data utilization [...], our experience suggests that such centers may not solve the desire for fast turn-around research or broaden access to those with unique perspectives. Synthetic data represent a promising approach for increasing easy access to secure data while simultaneously protecting the confidentiality of individuals."
>
> *Daniel Bonnéry, Yi Feng, Angela K. Henneberger, Tessa L. Johnson, Mark Lachowicz, Bess A. Rose, Terry Shaw, Laura M. Stapleton, Michael E. Woolley and Yating Zheng*

The GAN was devised in 2014 in Montreal, Canada by Goodfellow et al. at Université de Montreal. Two years before the start of the SDP, which must of been planned over a few years. It was too early for them to know about this obscure technique based on two neural networks competing against each other. Since then, Generative Adversarial Networks (Goodfellow et al., 2014) has inspired **23805** citations and algorithms capable of synthesizing data of impeccable similitude. We have surveyed a multitude of GAN algorithms built on the same basic idea of trial and error against an opponent that learns your faults. Despite the simple concept, we've seen that their range of application is wide, in general as well as in the healthcare domain. The variety of architectures and techniques we've seen reflect the heterogeneity of health data. Seemingly the difficulty of achieving stable learning with GANs in general delayed their application to OHD, while in medical imaging the development boomed much earlier sustained by the success of CNN in other fields. Notably, the innovation in the field of CNN has not slowed down after a few algorithms obtained excellent performance in image classification, but has deepened and branched out. Research concerning OHD seems to be gaining momentum rapidly. We saw thoughtfully engineered algorithms designed for the characteristics of OHD. Crucially however, pushing the research further will require a community effort to discover and defined metrics and standards upon which we can base objective assessment of models and SD. The challenges posed by OHD are nothing but encouragement for investigation and interpretation that can further our understanding of GANs, machine learning and human health. Undoubtedly the innovations made for OHD will find matches in other fields which may share the same data troubles.

Conflict of interest

The authors declare that there is no conflict of interest.

OHD-GAN Acronyms

1D-CAE 1-dimensional Convolutional AE. 11

AC-GAN Auxiliary Classifier GAN. 8, 30

ADS-GAN Anonymization through data synthesis using GAN. 9, 30

AnomiGAN GANs for anonymizing private medical data. 8, 30

CGAIN Categorical GAIN. 8, 30

CONAN *Co*plementary patte**r**n **A**augmentatio**n**. 8, 30

corGAN corGAN. 7, 11, 29, 30

CTGAN Conditional Tabular GAN. 8, 11, 30

cWGAN-GP Conditional WGAN-GP. 7, 30

CWR-GAN Cycle Wasserstein Regression GAN. 7, 13

DP-auto-GAN DP-auto-GAN. 8, 30

ehrGAN Electronic Health Record GAN. 6, 12

EMR-WGAN EMR Wassertein GAN. 9, 11, 30

GAIN Generative Adversarial Imputation Network. 7–9, 13, 32

GANITE Generative Adversarial Nets for inference of Individualized Treatment Effects. 7, 13

GcGAN CorrNN and T-wGAN. 8, 30

GluGAN Blood Glucose GAN. 8, 30

HealthGAN HealthGAN. 6, 8, 11, 13, 18, 20, 30

HGAN Heterogeneous GAN. 9, 11, 22, 23, 28, 30

MC-ARAE Multi-categorical ARAE. 7

MC-GumbelGAN Multi-categorical Gumbel-softmax GAN. 7

MC-medGAN Multi-categorical medGAN. 7, 9, 10, 19, 30

MC-WGAN-GP Multi-categorical WGAN with Gradient Penalty. 7, 10

MedBGAN Boundary-seeking medGAN. 7, 10, 11, 30

medGAN medGAN. 6–8, 10, 11, 13, 19, 21, 22, 30, 32

MedWGAN Wassertein medGAN. 7, 10, 30

PATE-GAN Private Aggregation of Teacher Ensembles (PATE) framework applied to GANs. 7, 30

PPWGAN WGAN for Temporal Point-processes. 6

RadialGAN RadialGAN. 7, 12, 21

RC-GAN Recurrent Convolutional GAN. 6, 11, 13

RGAN Recurrent GAN. 6, 11, 13

RMB Restricted Boltzmann Machine. 13

RSDGM Realistic Synthetic Dataset Generation Method. 9, 30

SC-GAN Sequentially Coupled GAN. 7, 13

SSL-GAN Semi-supervised Learning with a learned ehrGAN. 6, 7, 30

T-wGAN Wassertein T-GAN. 8, 32

WGAN-DP WGAN with DP. 8, 30

WGAN-GP WGAN with Gradient Penalty. 7–9, 11, 19, 30, 32

Glossary

0xCE-VAE 0xCE variational auto-encoder. 16, 26

AD Attribute Disclosure. 6, 9, 19

ADTEP Adversarial Deep Treatment Effect Prediction. 7, 13, 30

AE Autoencoder. 6–8, 10, 11, 13, 25, 32, 33

ARAE Adversarially regularized autoencoder. 7, 32

ARM Association Rule Mining. 18

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). 7

BN batch-normalization. 6, 9, 11

BP Backpropagation IT. 8

CAE Convolutional AE. 7

CGAN Conditional GAN. 6–9, 30

CNN Convolutional NN. 7, 8, 12, 23, 31

CorrNN Learns a common representation of two views, taking into account their correlation. See (Jolicoeur-Martineau, 2019; Su, 2018) text. 8, 32

COVID-19 coronavirus disease 2019. 1

CQM Clinical Quality Measure. 28

CRMB Conditional Restricted Boltzmann Machine. 5, 8, 30

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). 7, 12

Cycle-GAN Cycle-consistent GAN. 7, 12, 13, 29

D-TS discrete time-series. 6

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 (Strimbu and Tavel, 2010). 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." (Technology, 2020) text=digital bio-markers, first=digital bio-markers. 6

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. 6

DO Privacy preservation method. See (Yale et al., 2019a) based on (Dwork, 2008; Prasser et al., 2017). text. 8

DP Differential privacy. 2, 4, 5, 7–9, 18, 32, 33

DP-SGD Differential private stochastic gradient descent. 6, 8

DSM Domain Specific Measure. 16

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. 8

DWP Dimension-wise prediction. 18

ECG Electrocardiogram. 24

EEG Electroencephalogram. 24

EHR Electronic Health Record. 1, 4, 7, 9, 13, 16, 24, 28, 29

FFN Feed-forward Network. 11, 18

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. 8, 36–38

GAN Generative Adversarial Network. 1, 3–14, 16, 19–21, 23–29, 31–36, 38

GRU Gated Recurrent Unit. 8

Gumbel-GAN Gumbel-Softmax GAN. 7, 8

HexaGAN Six component GAN. 9

HI Health Informatics. 1, 2

HI-VAE Heterogeneous-Incomplete VAE. 9

ICU Intensive Care Unit. 13

InfoGAN Information Maximizing GAN. 26

IoT Internet of Things. 6, 7

IRL Inverse RL. 25

IT Iterative Imputation. 8, 33

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 (Haupt, 2019). text. 5, 7, 12, 13, 34

KLD Kullback-Leibler divergence. 16

KS Kolmogorov-Smirnov. 16

LN Layer normalisation. 9, 11

LR Logistic-regression. 13, 15

LSTM Long Short-term Memory. 6, 7, 13

MaR Given a dataset with missing entries , the missingness depends only on the observed variables (Yoon et al., 2018b). text. 5, 10, 13

MB-Avg Alternative to mini-batch discrimination that performs well on categorical features to cope with mode collapse, see (Choi et al., 2017a) text. 6

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. 7

MI See PD. text. 7–9, 19, 35, 37

MIDA Multiple Imputation Denoising Autoencoders. 9

MIMIC Openly available dataset of deidentified health data associated with 60,000 intensive care unit admissions. It includes demographics, vital signs, laboratory tests, medications, a range of data types from physiological time-series to free-text interpretation of radiology imaging (Johnson et al., 2016b) text. 27

MIWAE Missing data IWAE. 9

ML Machine Learning. 2, 21, 25, 27, 37

MMD Maximum Mean Discrepency. 16

mode collapse Mode collapse occurs when the GAN training procedure fails to converge, or converges to an undesirable local minima resulting in a lack of variety in the generated samples. text. 21, 22

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).. 7, 8

NN Neural Network. 3, 6–8, 34, 37

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. 8, 37

OHD Observational Health Data. 1–5, 8, 9, 20, 21, 24, 28, 31, 32

OHD-GAN GANs for Observation Health Data. 3, 4, 24

PartBB Similar to to the FullBB setting with the attacker having the additional knowledge about the latent input z. text. 8, 38

PATE Differential privacy method, best described by Papernot et al.: "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." (Papernot et al., 2017, 2018) text. 7

PCA Principal Component Analysis. 17, 22

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.; Jayaraman and Evans. text. 6, 9

PD Probability Distribution. 11, 16, 36

PL Difference of NN-AA on the test set and on the training set. See (Yale et al., 2019a). text. 8

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." (Ackerman, 2009) text. 6

PR Posterior Regularization. 25

PS Posterior Regularization. 25

RDP Renyi Differential Privacy. 8, 16

re-identification attack See MI text. 5

RF Random Forest. 15

RL Reinforcement learning. 25, 35

RMSE Root Mean-Squared Rrror. 18

RNN Recurrent NN. 6, 8

RR Reproduction rate. 19

RV-TS real-valued time-series. 6–8

SC shortcut connections. 6

SD Synthetic Data. 2–5, 7, 8, 14–16, 18, 19, 21, 27, 30, 31

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.. 8, 25

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. (Chapelle et al.) text. 5

SSL Semi-supervised learning. 7, 12, 13

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. 7, 12

SVM Support Vector Machine. 15

T-GAN Training technique to stabilise training. Allows the introduction of real sample information into the process of training the the generator. See (Jolicoeur-Martineau, 2019; Su, 2018) text. 8, 33

t-SNE The t-Distributed Stochastic Neighbor Embedding clustering algorithm is a nonlinear dimensionality reduction technique commonly applied to high-dimensional data. See Van Der Maaten and Hinton (2008). text. 6, 14

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. 8

TRTS Train on synthetic, test on real. 18

TSTR Train on real, test on synthetic. 18

VAE Variational AE. 8, 9, 13

VCD Variational contrastive divergence. 6, 12

VS Variable Splitting. 8, 13

WBA Similar to the PartBB and FullBB settings, but with the attacker having full knowledge of the generator internals, including gradient information.. 8, 37

WGAN Wassertein GAN. 6–10, 32, 33

References

Domenico Cucinotta and Maurizio Vanelli. Who declares covid-19 a pandemic. *Acta Bio Medica Atenei Parmensis*, 91(1):157–160, Mar. 2020. doi: 10.23750/abm.v91i1.9397. URL https://www.mattioli1885journals.com/index.php/actabiomedica/article/view/9397.

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, August 2016. doi: 10.1007/s00392-016-1025-6. URL https://doi.org/10.1007{%}2Fs00392-016-1025-6.

Hamed Abedtash, Mustafa Ascha, Mark Beno, Clair Blacketer, David Blatt, Brian Christian, Gino Cloft, Frank DeFalco, Sara Dempster, Jon Duke, Sergio Eslava, Clark Evans, Thomas Falconer, George Hripcsak, Vojtech Huser, Mark Khayter, Greg Klebanov, Kristin Kostka, Bob Lanese, Wanda Lattimore, Chun Li, David Madigan, Sindhoosha Malay, Harry Menegay, Akihiko Nishimura, Ellen Palmer, Nirav Patil, Jose Posada, Nicole Pratt, Dani Prieto-Alhambra, Christian Reich, Jenna Reps, Peter Rijnbeek, Patrick Ryan, Craig Sachson, Izzy Saridakis, Paola Saroufim, Martijn Schuemie, Sarah Seager, Anthony Sena, Sunah Song, Matt Spotnitz, Marc Suchard, Joel Swerdel, Devin Tian, Don Torok, Kees van Bochove, Mui Van Zandt, Erica Voss, Kristin Waite, Mike Warfe, Jamie Weaver, James Wiggins, Andrew Williams, Seng Chan You, , and and. *The Book of OHDSI*. Travis CI, 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, June 2020. doi: 10.7326/m19-0878. URL https://doi.org/10.7326%2Fm19-0878.

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, July 2020.

E Capobianco. Imprecise data and their impact on translational research in medicine. *Front Med (Lausanne)*, 7:82, 2020.

Fida Kamal Dankar and Khaled El Emam. The application of differential privacy to health data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, Edbt-icdt '12, page 158–166, New York, NY, USA, "30 " # mar 2012. Association for Computing Machinery. ISBN 9781450311434. doi: 10.1145/2320765.2320816. URL https://doi.org/10.1145/2320765.2320816.

Albert Cheu, Adam Smith, and Jonathan Ullman. Manipulation attacks in local differential privacy. "20 " # sep 2019. URL http://arxiv.org/abs/1909.09630.

Emiliano De Cristofaro. An overview of privacy in machine learning. "18 " # may 2020. URL http://arxiv.org/abs/2005.08679.

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.

Jean Louis Raisaro. *Privacy-Enhancing Technologies for Medical and Genomic Data: From Theory to Practice*. PhD thesis, Lausanne, EPFL, 2018. URL http://dx.doi.org/10.5075/EPFL-THESIS-8307.

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. July 2020. URL http://arxiv.org/abs/2007.10760.

Xinjian Luo, Yuncheng Wu, Xiaokui Xiao, and Beng Chin Ooi. Feature inference attack on model predictions in vertical federated learning. ”20 ” # oct 2020. URL http://arxiv.org/abs/2010.10152.

Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to federated learning: A survey. ”4 ” # mar 2020. URL http://arxiv.org/abs/2003.02133.

Ted Laderas, Nicole Vasilevsky, Bjorn Pederson, Melissa Haendel, Shannon McWeeney, and David A. Dorr. Teaching data science fundamentals through realistic synthetic clinical cardiovascular data. *bioRxiv*, page 232611, April 2018. doi: 10/ghfdmc. URL 10/ghfdmc.

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, June 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), June 2019. doi: 10.1162/99608f92.4a85af74. URL https://hdsr.mitpress.mit.edu/pub/9qsbf3hz.

Milad Yousefi, Moslem Yousefi, F S Fogliatto, R P M Ferreira, and J H Kim. Simulating the behavior of patients who leave a public hospital emergency department without being seen by a physician: a cellular automaton and agent-based framework. *Braz. J. Med. Biol. Res.*, 51(3):e6961, ”11 ” # jan 2018. ISSN 0100-879x, 1414-431x. doi: 10.1590/1414-431x20176961. URL http://dx.doi.org/10.1590/1414-431X20176961.

Anuraag R Kansal, Ali Tafazzoli, K Jack Ishak, and Stanmira Krotneva. Alzheimer’s disease Archimedes condition-event simulator: Development and validation. *Alzheimers. Dement.*, 4:76–88, ”16 ” # feb 2018. ISSN 1552-5260, 2352-8737. doi: 10.1016/j.trci.2018.01.001. URL http://dx.doi.org/10.1016/j.trci.2018.01.001.

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, August 2017. doi: 10.1093/jamia/ocx079. URL https://doi.org/10.1093%2Fjamia%2Focx079.

Daniel Bonnéry, Yi Feng, Angela K Henneberger, Tessa L Johnson, Mark Lachowicz, Bess A Rose, Terry Shaw, Laura M Stapleton, Michael E Woolley, and Yating Zheng. The Promise and Limitations of Synthetic Data as a Strategy to Expand Access to State-Level Multi-Agency Longitudinal Data. *J. Res. Educ. Eff.*, 12(4):616–647, ”2 ” # oct 2019. ISSN 1934-5747. doi: 10.1080/19345747.2019.1631421. URL https://doi.org/10.1080/19345747.2019.1631421.

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), March 2019a. doi: 10.1186/s12911-019-0793-0. URL https://doi.org/10.1186%2Fs12911-019-0793-0.

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, December 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. April 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. August 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, June 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, November 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, December 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, February 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), October 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, 2019a. doi: 10.1007/978-3-030-18579-4_3. URL https://doi.org/10.1007%2F978-3-030-18579-4%5F3.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. February 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, June 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, July 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, September 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)*, December 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 *Sashimimiccai*, 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%5F37.

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, 2019.

Ramiro Camino, Christian Hammerschmidt, and Radu State. Generating Multi-Categorical Samples with Generative Adversarial Networks. July 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, October 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, November 2019c. ISBN 9781728118673. doi: 10.1109/bibm47256.2019.8983215. URL https://doi.org/10.1109{%}2Fbibm47256.2019.8983215.

Ramiro D. Camino, Christian A. Hammerschmidt, and Radu State. Improving Missing Data Imputation with Deep Generative Models. February 2019.

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, April 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*, April 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, KDHECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020*, volume 2675 of {*CEUR*} *Workshop Proceedings*, pages 90–94. CEUR-WS.org, 2020a. 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.

Ramiro D Camino, Christian A Hammerschmidt, and Radu State. Working with Deep Generative Models and Tabular Data Imputation. *First Workshop on the Art of Learning with Missing Values (Artemiss)*, page 6, June 2020.

Uiwon Hwang, Dahuin Jung, and Sungroh Yoon. Hexagan: Generative adversarial nets for real world classification. *arXiv preprint arXiv:1902.09913*, 2019.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, pages 4413–4423, 2019.

Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107501.

Lovedeep Gondara and Ke Wang. Mida: Multiple imputation using denoising autoencoders. May 2017.

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, January 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 validation 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, August 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{%5F}10.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016*, pages 399–410. Institute of Electrical and Electronics Engineers Inc., December 2016. ISBN 9781509052066. 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, October 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, November 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, September 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.

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, October 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, February 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, December 2018. ISSN 14374331. URL http://orcid.org/0000-0001-5451-3010.

Wikipedia contributors. Jaccard index - wikipedia. URL https://en.wikipedia.org/wiki/Jaccard{%5F}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, April 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{%5F}p563.htmlhttps://doi.org/10.1142/9789811215636{%5F}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, April 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.

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.

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. December 2019.

Tianqing Zhu, Dayong Ye, Wei Wang, Wanlei Zhou, and Philip Yu. More than privacy: Applying differential privacy in key areas of artificial intelligence. *IEEE Trans. Knowl. Data Eng.*, pages 1–1, 2020b. ISSN 1041-4347, 1558-2191. doi: 10.1109/tkde.2020.3014246. URL https://ieeexplore.ieee.org/document/9158374/.

Xiao Chen, Thomas Navidi, Stefano Ermon, and Ram Rajagopal. Distributed generation of privacy preserving data with user customization. "20 " # apr 2019c. URL http://arxiv.org/abs/1904.09415.

N Li, T Li, and S Venkatasubramanian. Closeness: A new privacy measure for data publishing. *IEEE Trans. Knowl. Data Eng.*, 22(7):943–956, July 2010. ISSN 1041-4347. doi: 10.1109/tkde.2009.139. URL http://dx.doi.org/10.1109/TKDE.2009.139.

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.

Kasun Bandara, Hansika Hewamalage, Yuan-Hao Liu, Yanfei Kang, and Christoph Bergmeir. Improving the Accuracy of Global Forecasting Models using Time Series Data Augmentation. August 2020. URL http://arxiv.org/abs/2008.02663.

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, 2020.

Stuart McLennan, Leo Anthony Celi, and Alena Buyx. COVID-19: Putting the General Data Protection Regulation to the Test. *JMIR Public Health and Surveillance*, 6(2):e19279, April 2020a. ISSN 2369-2960. doi: 10.2196/19279. URL http://dx.doi.org/10.2196/19279https://www.ncbi.nlm.nih.gov/pubmed/32449686https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7265798https://publichealth.jmir.org/2020/2/e19279/AllPapers/M/McLennanetal.2020-COVID-19-PuttingtheGeneralDataProtecti.

S Latif, M Usman, S Manzoor, W Iqbal, J Qadir, G Tyson, I Castro, A Razi, M N Kamel Boulos, A Weller, and J Crowcroft. Leveraging data science to combat COVID-19: A comprehensive review. *IEEE Transactions on Artificial Intelligence*, pages 1–1, 2020. ISSN 2691-4581. doi: 10.1109/tai.2020.3020521. URL http://dx.doi.org/10.1109/TAI.2020.3020521.

Jason Walonoski, Sybil Klaus, Eldesia Granger, Dylan Hall, Andy Gregorowicz, Andy Gregorowicz, George Neyarapally, Abigail Watson, and Jeff Eastman. Synthea? Novel Coronavirus (COVID-19) Model and Synthetic Data Set. *Intelligence-Based Medicine*, page 100007, October 2020. ISSN 26665212. doi: 10.1016/j.ibmed.2020.100007. URL https://linkinghub.elsevier.com/retrieve/pii/S2666521220300077.

Jessamyn Dahmen and Diane Cook. SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors*, 19(5), 2019. ISSN 1424-8220. doi: 10.3390/s19051181. URL http://dx.doi.org/10.3390/s19051181https://www.ncbi.nlm.nih.gov/pubmed/30857130https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6427177http://www.mdpi.com/resolver?pii=s19051181https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6427177/AllPapers/D/DahmenandCook2019-SynSys-ASyntheticDataGenerationSystemforHealthcareApplications.pdf.

Ahmed Sedik, Abdullah M Iliyasu, Basma Abd El-Rahiem, Mohammed E Abdel Samea, Asmaa Abdel-Raheem, Mohamed Hammad, Jialiang Peng, Fathi E Abd El-Samie, and Ahmed A Abd El-Latif. Deploying machine and deep learning models for efficient Data-Augmented detection of COVID-19 infections. *Viruses*, 12(7):769, "16 " # jul 2020. ISSN 1999-4915. doi: 10.3390/v12070769. URL http://dx.doi.org/10.3390/v12070769.

Andres Ledesma, Mohammed Al-Musawi, and Hannu Nieminen. Health figures: An open source JavaScript

library for health data visualization. *arXiv [cs.HC]*, May 2016. doi: 10.1186/s12911-016-0275-6. URL http://dx.doi.org/10.1186/s12911-016-0275-6.

Jakob Nielse. 10 heuristics for user interface design: Article by jakob nielsen. URL https://www.nngroup.com/articles/ten-usability-heuristics/.

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.

Anne Marie Delaney, Eoin Brophy, and Tomas E Ward. Synthesis of Realistic ECG using Generative Adversarial Networks. September 2019.

Tomer Golany and Kira Radinsky. PGANs: Personalized Generative Adversarial Networks for ECG Synthesis to Improve Patient-Specific Deep ECG Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:557–564, July 2019. doi: 10.1609/aaai.v33i01.3301557. URL http://www.kiraradinsky.com/files/pgans-personalized-generative.pdfAllPapers/G/GolanyandRadinsky2019-PGANs-PersonalizedGenerativeAdversarialN...forECGSynthesistoImprovePatient-SpecificDeepECGClassification.pdf.

F. Ye, F. Zhu, Y. Fu, and B. Shen. ECG Generation With Sequence Generative Adversarial Nets Optimized by Policy Gradient. *IEEE Access*, 7:159369–159378, 2019. ISSN 2169-3536. doi: 10.1109/access.2019.2950383.

Pu Wang, Borui Hou, Siyu Shao, and Ruqiang Yan. ECG arrhythmias detection using auxiliary classifier generative adversarial network and residual network. *IEEE access : practical innovations, open solutions*, 7:100910–100922, 2019b.

Pratik Singh and Gayadhar Pradhan. A New ECG Denoising Framework Using Generative Adversarial Network. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2020. ISSN 1545-5963. doi: 10.1109/tcbb.2020.2976981. URL http://dx.doi.org/10.1109/TCBB.2020.2976981https://www.ncbi.nlm.nih.gov/pubmed/32142452https://doi.ieeecomputersociety.org/10.1109/TCBB.2020.2976981https://dx.doi.org/10.1109/TCBB.2020.2976981.

N K N Aznan, A Atapour-Abarghouei, and S Bonner. Simulating Brain Signals: Creating Synthetic EEG Data via Neural-Based Generative Models for Improved SSVEP Classification. *arXiv preprint arXiv*, 2019. URL https://arxiv.org/abs/1901.07429AllPapers/A/Aznanetal.2019-SimulatingBrainSignals-CreatingSyntheticEEGDataviaNeural-BasedGenerativeModelsforImproved-SSVEPClassification.pdf.

Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. June 2018.

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, March 2020.

Cheng He, Shihua Huang, Ran Cheng, Kay Chen Tan, and Yaochu Jin. Evolutionary multiobjective optimization driven by generative adversarial networks (GANs). *IEEE Trans Cybern*, "30 " # apr 2020. ISSN 2168-2275, 2168-2267. doi: 10.1109/tcyb.2020.2985081. URL http://dx.doi.org/10.1109/TCYB.2020.2985081.

Jacob Schrum, Jake Gutierrez, Vanessa Volz, Jialin Liu, Simon Lucas, and Sebastian Risi. Interactive evolution and exploration within latent Level-Design space of generative adversarial networks. "31 " # mar 2020. URL http://arxiv.org/abs/2004.00151.

J Mu, Y Zhou, S Cao, Y Zhang, and Z Liu. Enhanced Evolutionary Generative Adversarial Networks. In *2020 39th Chinese Control Conference (CCC)*, pages 7534–7539. Institute of Electrical and Electronics Engineers (IEEE), July 2020. ISBN 9789881563903. doi: 10.23919/ccc50068.2020.9188912. URL http://dx.doi.org/10.23919/CCC50068.2020.9188912.

Andrey Kurenkov. A brief history of neural nets and deep learning. *Skynet Today*, 2020.

Sajjad Abbasi, Mohsen Hajabdollahi, Nader Karimi, and Shadrokh Samavi. Modeling teacher-student techniques in deep neural networks for knowledge distillation. *arXiv preprint arXiv:1912.13179*, 2019.

Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization, 2016.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning, 2018.

Zhiting Hu, Zichao Yang, Russ R Salakhutdinov, LIANHUI Qin, Xiaodan Liang, Haoye Dong, and Eric P Xing. Deep generative models with learnable knowledge constraints. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10501–10512. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8250-deep-generative-models-with-learnable-knowledge-constraints.pdf.

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, June 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?, January 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, January 2016a. 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. January 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), September 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.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Nips'16, page 2180–2188, Red Hook, NY, USA, 2016b. Curran Associates Inc. ISBN 9781510838819.

Anat Reiner Benaim, Ronit Almog, Yuri Gorelik, Irit Hochberg, Laila Nassar, Tanya Mashiach, Mogher Khamaisi, Yael Lurie, Zaher S Azzam, Johad Khoury, Daniel Kurnik, and Rafael Beyar. Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. *JMIR Med Inform*, 8(2):e16492, "20 " # feb 2020. ISSN 2291-9694. doi: 10.2196/16492. URL http://dx.doi.org/10.2196/16492.

Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. Secure and robust machine learning for healthcare: A survey. *IEEE Rev. Biomed. Eng.*, Pp, "31 " # jul 2020. ISSN 1937-3333, 1941-1189. doi: 10.1109/rbme.2020.3013489. URL http://dx.doi.org/10.1109/RBME.2020.3013489.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, June 2019. ISSN 2052-4463. doi: 10/ghb4bs. URL 10/ghb4bs.

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, October 2013. doi: 10.1007/s10994-013-5409-9. URL https://doi.org/10.1007%2Fs10994-013-5409-9.

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, July 2002.

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). April 2020b. 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, February 2015. doi: 10.1016/j.jalz.2014.11.005. URL https://doi.org/10.1016%2Fj.jalz.2014.11.005.

U.S. Bureau of Labor Statistics. American Time Use Survey Home Page. URL https://www.bls.gov/tus/home.htm.

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), June 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 2016a. 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, February 2015. doi: 10.1055/s-0034-1396215. URL https://doi.org/10.1055%2Fs-0034-1396215.

National minimum dataset (hospital events). URL https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/collections/national-minimum-dataset-hospital-events.

Alexia Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard gan. In *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR, July 2019. URL http://arxiv.org/abs/1807.00734.

Jianlin Su. Training generative adversarial networks via turing test. October 2018. URL http://arxiv.org/abs/1810.10948.

Kyle Strimbu and Jorge A. Tavel. What are biomarkers?, November 2010. ISSN 1746630x. URL /pmc/articles/PMC3078627/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3078627/.

Pharmaceutical Technology. Digital biomarkers: Healthcare trends, 2020. URL https://www.pharmaceutical-technology.com/comment/digital-biomarkers-healthcare-trends/.

Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, pages 1–19. Springer Berlin Heidelberg, April 2008. doi: 10.1007/978-3-540-79228-4_1. URL https://link.springer.com/chapter/10.1007/978-3-540-79228-4{%5F}1.

Fabian Prasser, Johanna Eicher, Raffael Bild, Helmut Spengler, and Klaus A. Kuhn. A tool for optimizing de-identified health data for use in statistical classification. In *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, volume 2017-June, pages 169–174. Institute of Electrical and Electronics Engineers Inc., November 2017. ISBN 9781538617106. doi: 10.1109/cbms.2017.105.

Johannes Haupt. Causal machine learning: Individualized treatment effects and uplift modeling, 2019. URL https://johaupt.github.io/causal%20machine%20learning/Uplift%5FITE%5Fsummary.html.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, 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*, 3:160035, 2016b.

Nicolas Papernot, Ian Goodfellow, Martín Abadi, Kunal Talwar, and Úlfar Erlingsson. Semi-supervised knowledge transfer for deep learning from private training data. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, October 2017. URL http://arxiv.org/abs/1610.05755.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018. URL https://arxiv.org/abs/1802.08908.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proceedings of the 28th USENIX Security Symposium*, pages 1895–1912. USENIX Association, February 2019. ISBN 9781939133069. URL http://arxiv.org/abs/1902.08874.

Michael J. Ackerman. Personalized medicine. *The Journal of medical practice management : MPM*, 25(3):194–195, 2009. ISSN 87550229. URL https://www.genome.gov/genetics-glossary/Personalized-Medicine.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-supervised learning - the mit press. URL https://mitpress.mit.edu/books/semi-supervised-learninghttp://mitpress.mit.edu/books/semi-supervised-learning.

Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2625, 2008. ISSN 15324435. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.