Synthetic data generation for tabular health records: A systematic review

Synthetic Data Generation for Tabular Health Records: A Systematic Review

Mikel Hernandez[a,*], Gorka Epelde[a,b,*], Ane Alberdi[c], Rodrigo Cilla[a], Debbie Rankin[d]

[a] *Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009, Donostia-San Sebastián, Spain*

[b] *Biodonostia Health Research Institute, eHealth Group, Paseo Doctor Begiristain, s/n, 20014, San Sebastián, Spain*

[c] *Mondragon Unibertsitatea, Loramendi Kalea 4, 20500, Arrasate-Mondragón, Spain*

[d] *School of Computing, Engineering and Intelligent Systems, Ulster University, Derry-Londonderry, United Kingdom*

Abstract

Synthetic data generation (SDG) research has been ongoing for some time with promising results in different application domains, including healthcare, biometrics and energy consumption. The need for a robust SDG solution to capitalise on advances in Big Data and AI technology has never been greater to enable access to useful data while ensuring reasonable privacy protections. This paper presents a systematic review from the last 5 years (2016-2021) to analyse and report on recent approaches in synthetic tabular data generation (STDG) with a focus on the healthcare application context to preserve patient privacy, paying special attention to the contribution of Generative Adversarial Networks (GAN). In total 34 publications have been retrieved and analysed. A classification of approaches has been proposed and the performance of GAN-based approaches has been extensively analysed. From the systematic review it has been concluded that there is no universal method or metric to evaluate and benchmark the performance of various approaches and that further research is needed to improve the generalisability of GANs to find a model that works optimally across tabular healthcare data.

Keywords: synthetic data generation, generative adversarial networks, privacy preserving data, data sharing, healthcare, artificial intelligence

1. Introduction

The technological evolution of recent years, together with the digitalisation of traditionally manual processes, has enabled the timely and extensive collection, processing and analysis of data for decision making in many application fields. The availability of this information has motivated a paradigm shift from traditional knowledge and experience-based decision making toward evidence-based decision making, and the development of rapidly emerging applications in different sectors.

Once the initial challenge of preparing data for the reliable extraction of value is overcome [1], a security and privacy dilemma arises when dealing with scenarios that involve sharing data with third-parties, which has been generated within an organisation or by individuals. The ability to share such data with third-parties in a secure and privacy-preserving manner presents a range of opportunities. These include the collaborative

*Corresponding author

Email addresses: mhernandez@vicomtech.org (Mikel Hernandez), gepelde@vicomtech.org (Gorka Epelde), aalberdiar@mondragon.edu (Ane Alberdi), rcilla@vicomtech.org (Rodrigo Cilla), d.rankin1@ulster.ac.uk (Debbie Rankin)

development of complete and heterogeneous datasets to enable the creation of more precise artificial intelligence (AI) models; the ability to enable or even challenge the scientific community to research and improve AI models over private or sensitive data and application needs; and to enable AI development efforts to be extended beyond an organisation's internal research team. Thus far, the exploitation of opportunities that arise from data sharing have been impeded by a number of challenges. Whilst unique challenges can exist within specific application domains and their corresponding regulatory requirements (e.g. when considering sharing individual health data, sensitive commercial information, or biometric data) they can be generalised as challenges relating to the protection of privacy, intellectual property and security.

Together with traditional data anonymisation techniques for privacy-preserving data publishing (PPDP) that could allow for data pooling, in recent years, research has targeted distributed privacy-preserving data mining (aggregating distributed analytics results) and machine learning model (federated learning) training as a means to avoid data sharing [2, 3]. Moreover, methods of protecting the privacy of outsourced data by implementing encryption techniques have also been proposed[4]. In 2010 a survey on PPDP [5] discussed common privacy preservation models and their support for different types of attack, anonymisation techniques and information utility metrics. Anonymisation techniques seek to balance the trade-off between disclosure risk and data utility in the final published data, rendering a modified version of the original dataset in such a way that individuals are no longer identifiable [6, 7]. However, the utility of data anonymised using these methods is often adversely impacted and the data remains susceptible to disclosure [8].

A solution that can potentially overcome these limitations involves the generation of fully synthetic data (SD) as an alternative to real data. Synthetic data generation (SDG) is one of the most promising but as yet underexploited technologies for enabling PPDP and distributed privacy-preserving analytics. While not containing any data from the original set SD is generated from a model that fits to a real data set. Research on this technology has been ongoing for some time with promising results in different application domains, including healthcare [9], biometrics [10] and energy consumption [11], and the need for a robust solution to capitalise on advances in Big Data and AI technology has never been greater. Moreover, a recent publication reports cases of re-identification in anonymised individual-level data shared in the COVID-19 context, leading to a reduction of critical information sharing. This study proposes the use of synthetic tabular data generation (STDG) to enable access to useful information whilst ensuring privacy [11].

1.1. Motivation

The motivation for this systematic review is to thoroughly analyse and report on recent advances in STDG research and the utility of a range of methods, with a focus on the healthcare context. The potential positive impact of AI-enabled healthcare solutions is significant across a diverse range of application subdomains. AI solutions can improve patient experience, engagement, adherence and outcomes, and provide better clinical decision-making tools for diagnostics and treatments. AI also has the potential to significantly reduce costs and burden on health services [12]. Therefore, the development of such AI solutions is reliant on the availability of data, or potentially SD. Furthermore, this comprehensive review has also focused on the contribution of generative adversarial networks (GANs) for STDG. Since their inception in 2014 [13], GANs have garnered significant attention and are considered one of the most interesting developments in AI in recent years. GANs have excelled in the generation of synthetic image data [14]. Given this promising performance, the development of GANs for alternative data types, particularly tabular data, is currently a popular topic in the AI research community and as such this work has identified a substantial group of related publications in the last few years.

1.2. Research Questions (RQ)

Since the aim of this systematic review is to analyse the approaches used for STDG in healthcare paying special attention to the contribution of GAN approaches, the RQ that have guided this review are as follows:

- *RQ1:* What approaches exist for generating synthetic tabular data in healthcare? How can these methods be classified?
- *RQ2:* Which of these approaches are based on GANs? What are their characteristics and/or distinctions?
- *RQ3:* What is the performance of these approaches in creating data that are usable, private and similar but not identical to real data?

1.3. Contributions

In this work a comprehensive study is conducted to investigate the different STDG approaches that have been used in healthcare to preserve privacy. Specifically, we carry out a systematic review on STDG in healthcare, with a specific focus on the contribution of GAN technologies to this research area. To the best of our knowledge, this work is the first attempt to review STDG approaches in healthcare to preserve privacy with a focus on GAN technology. However, Georges-Filteau and Elisa [15] reviewed STDG in healthcare using GANs with a broader focus of STDG in healthcare (data augmentation, privacy preservation, etc.), which provides a more general and less privacy-specific overview of STDG in healthcare with GAN technology.

1.4. Article outline

The remainder of the paper is organised as follows: In Section 2 concepts related to SDG and healthcare tabular data are presented. The methodology employed to develop the systematic review is described in Section 3. In Section 4 the publications retrieved by applying the previously explained methodology are presented. Section 5 answers RQ1 describing the used STDG approaches used in the publications analysed and providing a complete taxonomy of them. Section 6 answers RQ2 and RQ3, giving an overview of the GAN-based STDG approaches, explaining the evaluation methods used to evaluate them and benchmarking the GAN-based approaches. Finally, Section 7 provides a general interpretation of the systematic review results and the research gaps identified are discussed. Section 8 concludes the work developed.

2. Background

2.1. Synthetic Data: Use Cases

SD is artificial data generated by a model trained or built to replicate real data (RD) based on its distributions (i.e., shape and variance) and structure (i.e., correlations among the attributes) [16]. SD has two main use cases, (i) data augmentation: to balance datasets or to complement available data before training a ML model; (ii) privacy-preservation: to enable secure and private sharing of sensitive data (the goal of this analysis). Moreover, SD can potentially be used for drawing statistical conclusions or training ML models, while preventing the divulgence of sensitive data.

SDG is among the most promising privacy preservation techniques, as it does not contain data from the original set and leads to lower information loss in the resultant synthetic dataset. However, prior to adoption, SDG techniques need to be assessed in terms of privacy (personal data disclosure risk), resemblance (how well the SD represents the real data), utility (usability of statistical conclusions drawn from SD or the results from SD trained ML models) and performance dimensions (footprint, generation time and computational resources).

2.2. Synthetic Tabular Data Generation in healthcare

Within healthcare, SDG has been researched for different modalities including biomedical signals [17], medical images [18], electronic health records (EHR) free-text content [19], time-series smart-home activity data [20] and EHR tabular data [21]. In this systematic review we focus on STDG, as tabular data is the largest exponent of structured data. Despite recent trends that combine non-structured data with structured health data [22], the most common approach for developing ML models for healthcare decision making has been through structured data and the structuring of unstructured data through coding or feature extraction algorithms. Therefore, tabular health data potentially offers the most valuable opportunities in the development of AI-based health care systems and progress is lacking as data subjects in structured data are easier to identify compared to other modalities [12].

Tabular or structured healthcare-related data stored in EHR, clinical trials or labs contain vast and diverse amounts of patient-related data, with linkages across different data sets. Normally, each row of a tabular healthcare dataset correspond to one record of data that include; (1) explicit identifiers that uniquely identify the record (e.g. id and name), (2) quasi-identifiers that cannot uniquely identify the record and capture time-independent descriptive patient data (e.g. date of birth and genre and demographics data), and (3) sensitive attributes that predominantly represent longitudinal data consisting of a set of medical events at different time-steps (e.g. diagnosis, lab test results and prescription data) [23]. More recently, enabled by sensor technology, we can access vital-sign data in (more frequent) time-series format. Moreover, tabular data variables can consist of various data types including categorical (including binary), ordinal, numerical, and dates. For the analysis of targeted techniques, it is necessary to consider this diversity.

2.3. Privacy-preserving data in healthcare

According to Donaldson and Lohr, privacy is defined as the act of being kept away from public view, but with no pejorative overtones [24]. Newer studies state that healthcare privacy can be defined by contextual rules about how information or personal data can flow to maintain the security and confidentiality of patient records. These rules depend on the actors involved, the process by which information is accessed, the frequency of the access and the purpose of that access [25]. When they are contravened an "*invasion of privacy*" or "*violation of privacy*" occurs causing a disclosure of personal information that users intend to keep private from an entity which is not authorised to obtain the data [24, 25, 23].

Regarding "*privacy violation*", Chong identified some privacy threats including privacy disclosures and attack models [23]. Privacy disclosures include: (1) identity disclosure or reidentification (when the true identity of a patient is revealed by an adversary), (2) attribute disclosure (when one or more sensitive attributes of a patient is revealed by an adversary) and (3) membership disclosure (when the existence of a patient in the published data is successfully inferred by an adversary). Having some previous knowledge about the data (e.g., published dataset, or quasi-identifiers of a patient) an adversary can use linkage attacks, homogeneity attacks, background knowledge attacks, skewness attacks and similarity attacks to successfully perform a disclosure of privacy in terms of identity, attribute or membership.

In our work, we define private healthcare data as data which: (1) maintains patterns of real data, (2) is similar but not identical to real data, (3) does not contain sensitive information of real patients and (4) can be used for conclusion drawing and new knowledge generation. Thus, to make advances in the use of AI technologies, which require a high volume of data, PPDP methods need to be analysed in order to avoid privacy violation. PPDP-related studies in healthcare are focused on the anonymisation of data and the analysis of the trade-off between data utility and privacy. Examples include the survey studies authored by Chong [23], Fung et al. [5], Tran and Hu [26] and Wang et al. [27], in which they reviewed privacy preservation models and techniques, measuring information utility with various metrics, and privacy under

different types of attack, however none of these studies consider STDG as a privacy preservation method for healthcare-related data. Although Chester et al. [28] proposed some metrics and methods to balance utility and fairness against privacy in medical data, classical techniques for PPDP do not offer results comparable to STDG.

STDG has been analysed and reviewed as an alternative for high-quality PPDP with satisfactory results in some studies. On the one hand, Azizi et al. [29] validated the use of SD replicating the analysis from a study published on a real dataset showing that synthetic data can be used as a reasonable proxy for real clinical trial datasets. On the other hand, El Emam et al. [30] developed and applied a methodology for evaluating the identity disclosure risks of fully synthetic data using a COVID-19 cases database from Canada and concluding that STDG can reduce meaningful identity disclosure risks considerably. Furthermore, Dankar and Ibrahim [31] investigated the effect of various STDG settings on the quality of the generated data to provide the best strategies to follow when generating and using synthetic data.

3. Systematic Review Process

To conduct the systematic review we have defined a six-step methodology, starting from the proposals made by Khan et al [32] and Uman [33]. First, a bibliographic search strategy is established. Second, the bibliographic search strategy is performed and relevant publications are identified according to the defined inclusion/exclusion criteria. Third, the information or data is extracted through the publication reading and synthesising process. Fourth, the identified models or approaches are evaluated, compared and benchmarked according to metrics relevant to the review topic under study. In the sixth and final step, results obtained from the data information extraction have been analysed and interpreted to summarise the state of the research and draft new investigation lines. The remainder of this section covers the steps of this methodology.

3.1. Search Strategy

To define the search strategy in a reproducible and unambiguous manner, it has first been necessary to define the following items: (1) The search engines to be used to perform the scientific database search; (2) the search limits set to constrain the retrieval of results; and (3) the inclusion and exclusion criteria to retain only the most relevant publications.

3.1.1. Search engines

The selected search engines for the current systematic review have been Engineering Village[1], Scopus[2], Web of Science[3] and PubMed[4]. The combination of these engines offers a wide coverage of general scientific publications. Additional publications identified from reference lists of the review's main articles and alerts from Google Scholar have also been considered as candidates for this review.

3.1.2. Search Limits

The following search limits were applied:

- *Search terms*

[1] https://www.engineeringvillage.com/home.url
[2] https://www.scopus.com/
[3] https://apps.webofknowledge.com/
[4] https://pubmed.ncbi.nlm.nih.gov/

A preliminary analysis of the state of the art in STDG in healthcare enabled the identification of relevant keywords and search terms covering several aspects of the topic, such as the nature of synthetic data, data generation issues, privacy preserving data or different data types (i.e. tabular data, patient records, medical data, time-series healthcare data).

Roots of the identified relevant terms were used for generalisation purposes. In addition, these search terms were strategically classified as keywords to be searched only in the title field or in any of the following fields: title, abstract, keywords or topic. As a result, two search strings were defined using logical operations: (1) *(synthe* OR generat* OR privacy) AND (tabular OR data OR record OR time-series OR sample)* to be searched on the title field and (2) *("synthetic data" OR "generative model") AND (patient* OR medic* OR health* OR clinic*)* to be searched in the title, abstract, keywords and/or topic fields.

- *Publications' date range*

To find the most groundbreaking studies, we limited our search to publications from the last five years, starting from 1st January 2016 to 17th May 2021 (date of the bibliographic search).

- *Type of publications*

We limited the search to peer-reviewed conference and journal articles written in the English language. These kinds of publications are considered to provide a good view of accepted and validated methodologies and knowledge.

3.1.3. Inclusion and Exclusion criteria

After gathering the publications from the search engines, the most relevant studies have been selected based on the defined inclusion and exclusion criteria.

- *Inclusion criteria*

The publications that were included in this systematic review met the requirement of working with tabular healthcare data. Additionally, they fulfilled at least one of the following conditions:

1. They implement an approach or method to generate tabular synthetic data.
2. They compare various synthetic data generation approaches or methods.
3. They evaluate the synthetic data generated by one or more approaches.

- *Exclusion criteria*

In order to reject all possible irrelevant publications retrieved, additional exclusion criteria were added to those implicitly defined by the search strategy (section 3.1) and inclusion criteria. As a result, the publications that were excluded from this systematic review fulfilled at least one of the following conditions:

1. The publication does not deal with synthetic data generation.
2. Some data type other than tabular has been considered (image, text, videos, signals, etc).
3. The publication is a short version of another retrieved publication.
4. The publication has not been peer-reviewed.
5. The publication is not written in English.

3.2. Data extraction and Synthesis of results

Finally, the information required to answer the RQ from the retrieved publications was extracted and synthesised for each retrieved publication. First, in response to RQ1, the suggested and/or utilised STDG approaches were identified. For RQ2, the characteristics of the proposed GAN models have been synthesised from publications where such an approach was used. Lastly, to answer RQ3, information about the

selected evaluation metrics and the results obtained were extracted. Some additional information was considered to be pertinent and therefore extracted, including the purpose of the study (e.g. data augmentation, preservation of privacy) and the type of data used.

It is worth mentioning that each study uses different evaluation metrics to evaluate their method's performance in terms of resemblance, utility, and privacy, and it was therefore not possible to use a single universal metric to compare them. To overcome this issue, we categorised the models' performance as '*Excellent*", "*Good*" and "*Poor*" according to the results reported in comparison with other methods in the same study enabling a per-paper-basis comparison. A more detailed description of this categorisation process can be found in section 6.2.

4. Overview of the Selected Publications

In this section, an overview of the systematic review process execution results and the publications retrieved is presented. Figure 1 illustrates the step-by-step results of the systematic review. 346 publications were collected from the search engines and 20 publications were manually retrieved by references or alerts, giving a total of 366 publications. Duplicated studies (190) from the merged list of all databases (including the publications added manually) were excluded. At this point, 176 publications remained for the screening process, of which 92 were excluded after reviewing titles and abstracts. In the eligibility step, 84 publications remained for full-text assessment, of which 50 publications were excluded: 45 of them for not meeting the inclusion criteria, 2 for not being written in English, and 3 for being short versions of other included publications. After applying this selection process, 34 publications were retained for the systematic review. These results were obtained from a bibliographic search executed on 17th May 2021.

In Table 1 a brief description of the 34 selected publications is provided, presenting the purpose of the study, the data used and study types, the applied or proposed STDG approaches and the evaluation methods used. The results show that the purpose of STDG in the majority of the studies analysed is preserving privacy to enable medical data sharing, which is the motivation of this systematic review. Additionally, a high variety in the context of the data sources is observed, with EHRs and ICD9 codes the most repeated sources. In many of the studies, mainly categorical and time-series data is synthesised. Regarding the STDG approaches and evaluation metrics, there is wide diversity within the publications. There are not many repeated approaches, and the evaluation methods are varied. Furthermore, not all authors evaluate the three defined dimensions of SD (resemblance, utility and privacy). Some evaluate one or two of them, and others all of them. Privacy is the least evaluated dimension of synthetic data in the identified publications.

Table 1: Brief description of the included publications

Publication	Purpose	Data type	Study type	STDG Approaches	Evaluation methods
McLachlan 2016 [34]	Preserve privacy	-	Midwifery EHR	CoMSER	(U)Consultation with clinical experts
Che 2017 [35]	Augment data	Categorical Time-series	ICD9 diagnoses and medications (a)Congestive heart failure (b)Diabetes	ehrGAN	(U)Augment data for ML model training (R)Compare data length distributions (R)Compare frequency of top 100 features

Table 1: Continued on next page

Table 1 (cont)

Publication	Purpose	Data type	Study type	STDG Approaches	Evaluation methods
Choi 2018 [36]	Preserve privacy	Binary	(a)ICU patients (b)Heart failure	GAN GANp GANpd GANpa medGAN* RN IS DBM VAE	(R)Dimensional probability (R)Dimensional prediction (R)Consultation with clinical experts (P)Identity disclosure (P)Attribute disclosure
Park 2018 [37]	Preserve privacy	Numerical Categorical Time-series	(a)LA City worker records (b)Records of personal information (c)Records of medical information	table-GAN* DCGAN Condensation method k-anonymity + t-closeness DP + δ-disclosure Micro-aggregation Post-randomisation	(U)Compare cumulative distributions (U)TRTR and TSTR (P)DCR (P)Membership attack
Walonoski 2018 [38]	Preserve privacy	Categorical Numerical	Diabetic diseases	Synthea	(U)Compare average values with real average values
Norgaard 2018 [39]	Preserve privacy	Numerical Time-series	Daily and Sports activities	Supervised GAN	(R)STS, RTS and RTR (P)Max-RTS (U)ML models: TSTR
Wu 2018 [40]	Preserve privacy	Numerical	USA Census data	CMEM	(R)STS, RTS and RTR (P)Max-RTS (U)ML models: TSTR
Zare and Wojtusiak 2018 [41]	Propose SD evaluation metric	Binary	-	Logistic Regression	(R)WIE
Vaidya 2018 [42]	Preserve privacy	Numerical Categorical	(a)Breast Cancer (b)Parkinson's Telemonitoring (c)Diabetes	RDT	(U)ML models: TRTR and TSTR
McLachlan 2019 [43]	Preserve privacy	Numerical Categorical	Labour and Birth EHR	ATEN Framework	(R)Statistical values comparison (R)Survey with clinical experts
Wang 2019 (1) [44]	Preserve privacy	Categorical Time-series	(a)Sepsis-3 (b)Diabetes	SeqGAN C-RNN-GAN RCGAN SC-GAN	(R)Dimension-wise probability (R)Pairwise pearson correlation (R)Consultation with clinical experts (U)ML models: TRTR and TSTR (U)Augment data for ML models training
Jackson and Lussetti 2019 [45]	Preserve privacy	Categorical	ICU Patients	Extended medGAN	(R)Compare most common values
Dahmen and Cook 2019 [20]	Augment data	Numerical Time-series	Smart Home environment	SynSys	(R)Compare one week of real data with synthetic data (R)Euclidian distance (R) DTW (U)Augment data for ML models training
Beaulieu-Jones 2019 [46]	Preserve privacy	Numerical Categorical	(a)SPRINT (b)ICU Patients	AC-GAN AC-GAN + DP	(R)Consultation with clinical experts (U)ML models: TRTR and TSTR (P)Formulation of DP
Wang 2019 (2) [47]	Preserve privacy	Categorical Time-series	UK Primary Care - CVD	BN	(R)Distance between real and synthetic values (probability and distribution) (R)KS test (R)PCA, NMDS, Correlation matrix (P)DBSCAN

Table 1: Continued on next page

Table 1 (cont)

Publication	Purpose	Data type	Study type	STDG Approaches	Evaluation methods
Abay 2019 [48]	Preserve privacy	Numerical Categorical	(a)CMC (b)Mamographic Mass (c)Diabetes (d)Breast Cancer	PrivateSVM PrivBayes DP-EM DP-VAE DP-SYN	(R)Compare probability distributions (U)ML models: TSTR
Chin-Cheong 2019 [49]	Preserve privacy	Numerical Categorical	EHR from New Zealand's health care system	WGAN DP-WGAN	(R)Compare distribution of variables (U)ML models: TRTR and TSTR
Yang 2019 [50]	Augment data	Categorical	ICD9 codes: EHR from paediatric department of a hospital	WGAN T-WGAN medGAN ehrGAN CorrGAN GcGAN	(R)Compare the mean and sd (R)Compare the frequency of data features (U)Augment data for ML models training
Baowaly 2019 [51]	Preserve privacy	Numerical Binary Time-series	(a)ICU Patients (b)Taiwan NHIRD	medGAN medWGAN medBGAN	(R)Compare dimensional probability (R)Compare dimensional average (R)Dimensional KS test (U)ML models: TRTR and TSTR
Yale 2020 (1) [21]	Preserve privacy	Discrete Categorical	ICU Patients (a)Impact of race on 30-day mortality (b)Mortality of elderly patients (c)Mortality of patients with acute renal injury	medGAN HealthGAN GM PW ANM DP CRD	(R)Compare dimensional probability (R)Nearest neighbor AA (R)Resemblance loss (R)PCA, Histograms and Correlation matrix (P)Privacy loss
Yale 2020 (2) [52]	Preserve privacy	Discrete Categorical	Co-occurring conditions in ASD	HealthGAN	(R)Compare dimensional probability (R)Nearest neighbor AA (R)Resemblance loss (U)Cox regression and k-means clustering (P)Privacy loss (P)Membership inferences attact scenario
Dash 2020 [53]	Preserve privacy	Numerical Categorical Binary Time-series	(a)Sleep patterns of people over a 30-hour period (b)ICU patients: hospital mortality and phenotype classification	HealthGAN TimeGAN	(R)Compare average trends (R)Welsch t-test (U)ML models: TRTR, TRTS, TSTS and TSTR
Rashidian 2020 [54]	Preserve privacy	Binary Categorical Numerical	Inpatient encounters with elderly patients ($\geq$ 18)	SmoothGAN cGAN AC-GAN WGAN WGAN-GP	(R)MAE for means and sd (R)Compare Pearson correlation coefficients (U)ML models: TRTR, and TSTR (P)MMD
Yoon 2020 [55]	Preserve privacy	Binary Categorical Numerical	(a)MAGGIC (b)UNOS	ADS-GAN PATE-GAN DP-GAN medGAN WGAN-GP	(R)Student t-test and Chi-squared test (U)ML models: TSTR (P)JSD and Wassestein distance
Rankin 2020 [9]	Preserve privacy	Binary Categorical Numerical	19 open health care datasets	CART Parametric (LR, LOG REG. and Polytonous LOG. REG.) BN	(R)Compare multivariate relationships (U)ML models: TRTR and TSTR (P)SDC metrics

Table 1: Continued on next page

Table 1 (cont)

Publication	Purpose	Data type	Study type	STDG Approaches	Evaluation methods
Lee 2020 [56]	Preserve privacy	Categorical Time-series	ICD9 codes (a)EHR of ICU patients (b)Outpatients of UT Physicians	DAAE medGAN VAE VAE-GAN WAE ARAE	(R)ML model to classify records in real or synthetic (R)Compare plausibility scores with clinical experts (R)DBSCAN (U)ML models: TSTR (P)Differential privacy cost
Fowler 2020 [57]	Augment data	Numerical Categorical	Mammography studies from Moffitt Cancer Center	MKDE	(R)ML model to classify records in real or synthetic (R)Kernel two-sample test based on MMD (R)Compare the covariance matrices (R)PCA
Goncalves 2020 [58]	Preserve privacy	Categorical	SEER program (a)Breast cancer (b)Respiratory cancer (c)Non-solid cancer	MPoM CLGP MC-medGAN MICE-DT	(R)KL divergence (R)PCD (U)Log-cluster (U) Support coverage (U)Cross-classification (P)Identity disclosure (P)Attribute disclosure
Hyun 2020 [59]	Augment data	Numerical Categorical Time-Series	ICU Data: Diabetes	Prophet	(R)KDE Comparison (R)KL Divergence (R)PCD
Wang 2020 [60]	Preserve privacy	Numerical Time-Series	ICU Data from US	DP-GAN PART-GAN	(R)Statistics and cumulative distributions comparison (P)Inception Score (P)Euclidean distances
Koivu 2020 [61]	Augment data	Numerical Categorical	Early stillbirth prediction	actGAN	(U)ML models: Augment data
Tucker 2020 [62]	Preserve privacy	Numerical Categorical	UK Clinical Practice Research Datalink - CVD	BN	(R)Statistical Tests (R)KL divergence (R)MMD
Wang 2021 [63]	Preserve privacy	Numerical Categorical	(a)Indian Liver Patient Disease Dataset (b)CVD prediction	BN GM	(R)KS Test (R)Correlation matrices distances (U) ML models: Augment data (P)DBSCAN
Zhang 2021 [64]	Preserve privacy	Categorical Time-Series	ICD9 Codes and Demographics	SynTEG	(R)Correlations between temporal features (U)Future diagnosis forecasting (P)Membership inference (P)Attribute disclosure

Evaluation methods: (R), Resemblance evaluation; (U), Utility evaluation; (P), Privacy evaluation.

5. Synthetic Tabular Data Generation Approaches in Healthcare

This section presents the STDG approaches found in the publications from this systematic review. Mainly, machine learning (ML) and deep learning (DL) based approaches have been used to synthesise medical data, while baseline methods and statistical and probabilistic models are used less often. In a small number of cases the authors propose a framework or process based on different algorithms and/or models. Considering this and answering RQ1, all the approaches used for STDG can be classified in three main groups: classical approaches, deep learning approaches and others. An extended classification of the approaches is exposed in Figure 2.

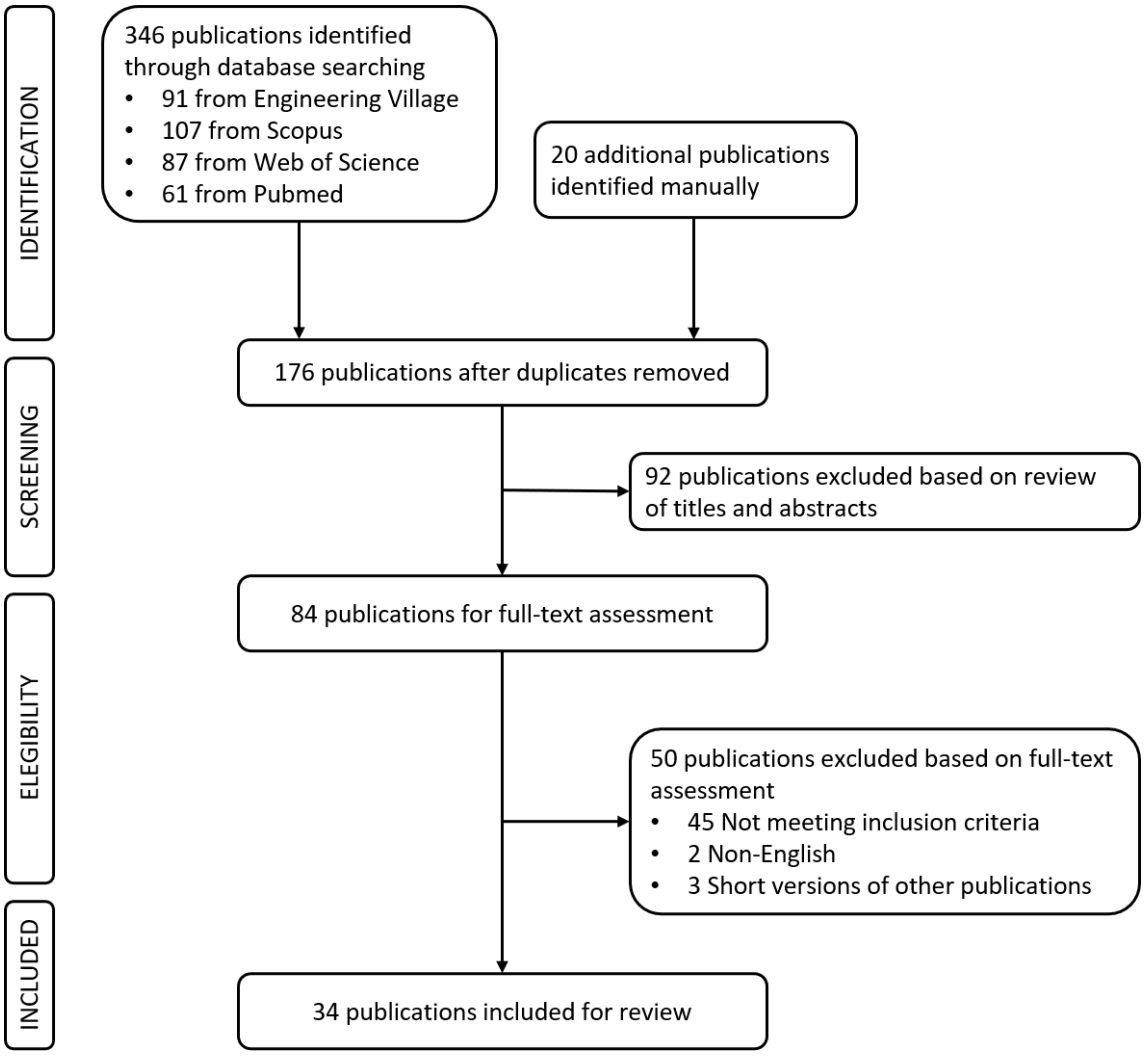


Figure 1: Flow diagram of the publications' selection process

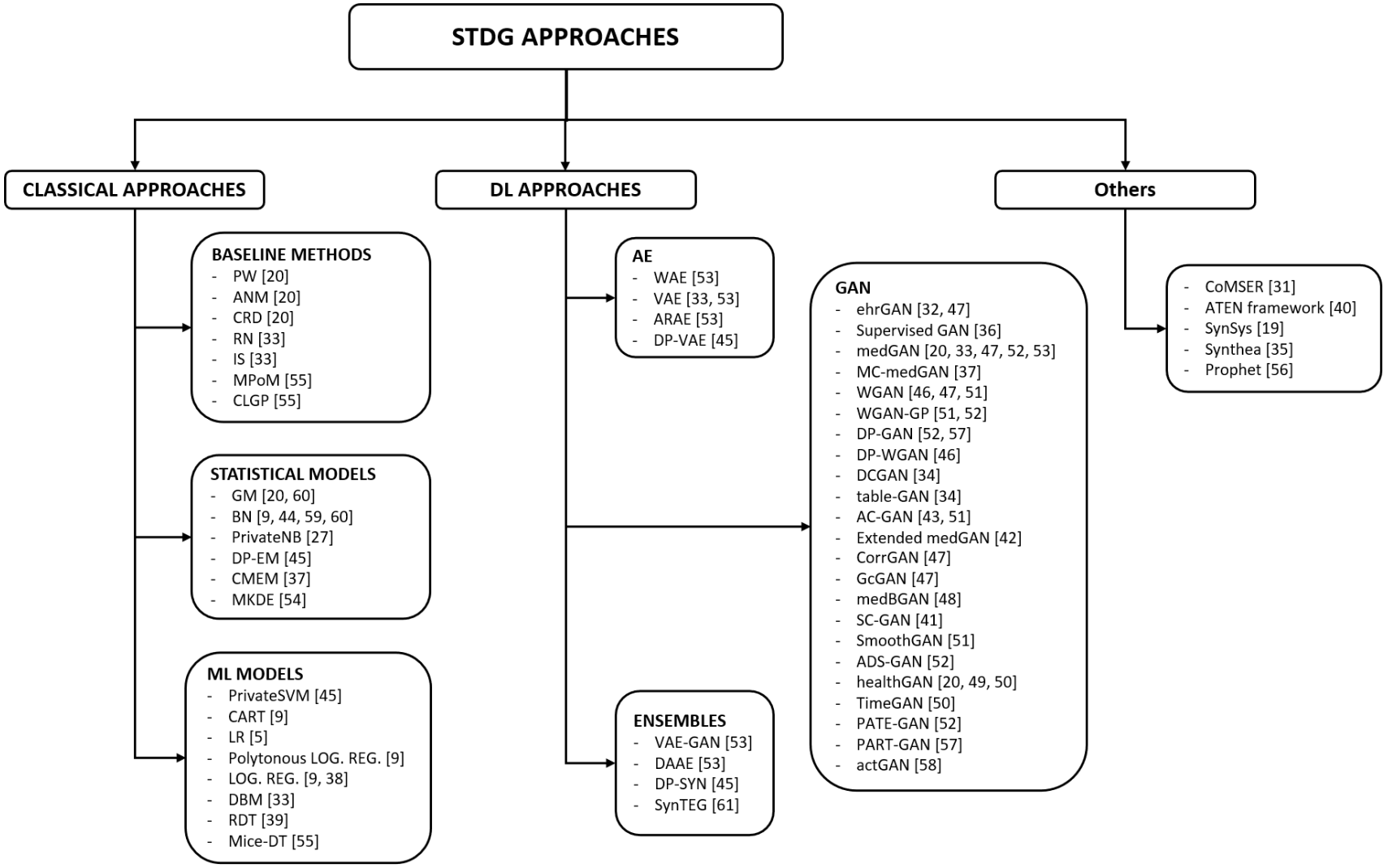


Figure 2: Proposed classification of the STDG approaches found in the included publications

5.1. Classical Approaches

Among the classical approaches, baseline methods, statistical and probabilistic models and ML models can be distinguished.

5.1.1. Baseline Methods

Baseline methods are related to anonymisation techniques and/or methods that do not use data modelling to generate synthetic data. Generally, these techniques are based on replacing values, deleting sensitive attributes and adding noise to the data. However, they do not offer usable and private results. As Yale suggested in [65] these methods try to memorise and summarise each feature and the relationships between them, to then use these results to generate the SD. For this reason they do not offer a good balance between resemblance, utility and privacy.

5.1.2. Statistical Models

Approaches categorised in this group synthesise the data using statistical and probabilistic models that attempt to simulate the real data. They usually capture correlation structures between attributes in the real data (RD) and samples are drawn from the model to generate the synthetic data. These approaches have been used to generate categorical and numerical data, and few of them have been tested on time-series data.

5.1.3. Machine Learning (ML) models

Supervised ML models have also been used for STDG. With RD being the input of these models, they learn to predict new data records that are very similar to the original data. Normally the data is synthesised sequentially, i.e., the model is used to predict the next time-steps or sequences of patients and those are considered as the synthetic data. Some of these models have been used for numerical data (Linear Regression) and others for categorical (Polytomous Logistic Regression) or binary data (Logistic Regression). The rest of the models have been used to synthesise mixed data types at the same time.

5.2. Deep Learning (DL) Approaches

The approaches that have been considered in the DL group are those that are composed of neural networks. Within this group, Autoencoders (AE), GANs and Ensembles can be distinguished.

5.2.1. Autoencoders (AE)

Autoencoders are a type of unsupervised neural network that learn how to reconstruct data given an encoded representation of the RD. Originally, they are composed of an encoder and a decoder. The encoder efficiently compresses the data, while the decoder decompresses this data in a close representation of the encoded version. These methods have been used to generate a wide variety of biomedical data types, but predominantly categorical and binary data.

5.2.2. Generative Adversarial Networks (GANs)

A GAN is a specific type of DL model that principally consists of two neural networks (generator and discriminator). These two neural networks learn to generate high quality SD by an adversarial training process. Variants of the original GAN and other approaches are repeatedly used by different authors, in some cases presenting improvements, tuning some hyperparameters or adding new features. A more detailed description of those methods is presented in section 6.

5.2.3. Ensembles

Ensemble methods have been defined as those methods in which two different types of DL models are used to generate synthetic data. The reported results suggest that these techniques can not outperform the previously mentioned methods based on a single DL model.

5.3. Other Approaches

Finally, techniques that do not fit within previous categories have been grouped in this category. These include methods, procedures or frameworks that are composed of several steps or modules, and are presented below.

5.3.1. Content Modelling for Synthetic E-Health Records (CoMSER)

The Content Modelling for Synthetic E-Health Records (CoMSER) method was proposed by McLachlan et al. in 2016 [34] to generate synthetic EHRs based on the use of publicly available Health Information Statistics (HIS) and the knowledge collected from experienced clinicians as a two-step procedure.

5.3.2. Aten Framework

The Aten framework was proposed by McLachlan et al. in 2019 [43] to generate synthetic Labour and Birth EHRs whilst characterizing and validating its realism by gathering the necessary knowledge, identifying realistic properties from real data and validating the realism of SD.

5.3.3. SynSys

SynSys was proposed by Dahmen and Cook in 2019 to generate realistic synthetic smart home sensor data by means of Hidden Markov models (HMM) [20]. It creates temporary sequences of a variety of daily activities in a smart home environment training a different model for each activity.

5.3.4. Synthea

Synthea is another approach for generating synthetic EHRs. It was proposed by Walonoski et al. in 2018 [38] as an open-source synthetic health simulator that simulates synthetic patients from birth to death in JSON format via different pre-defined modules (diabetes, cancer, infections, treatments, etc). Chen et al. [66] validated the use of this system to generate a cohort to analyse various clinical measures and Walonoski et al. [67] used it to create synthetic data related to the COVID-19 pandemic.

5.3.5. Prophet

Prophet is a procedure used by Hyun et al. [59] for forecasting time-series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It is robust to missing data and shifting trends, and typically handles outliers well. Additionally, it can sufficiently express medical changes with its components, analyse trends and effectively predict values of a target parameter.

6. GAN-based Synthetic Tabular Data Generation Approaches in Healthcare

This section aims to answer RQ2 and RQ3, by presenting the contribution and performance of GANs in STDG tasks. As Alqahtani et al. defined in [14] the basic GAN architecture is composed of two types of neural networks: the generator (G) and the discriminator (D). While G is used to generate synthetic samples using a noise vector as input, D determines whether the samples generated by G belong to a real distribution or not. Through an adversarial training process, G learns to generate increasingly realistic samples using the feedback from D. A number of GAN variants are analysed and explained below and details of how they are evaluated are given.

6.1. Description of the GAN-based approaches

This section discusses the GAN-based approaches found in this systematic review for STDG to answer RQ2. Some other approaches have been omitted because their initial purpose is not to generate tabular data, however some of them have been used in the publications to make comparisons with the proposed method.

6.1.1. ehrGAN

Che et al. adapted the basic GAN model for generating realistic EHR in order to boost DL prediction models, proposing ehrGAN [35]. In this work, sequential ICD9 diagnosis and medication codes of patients were used. The D is composed of a unique layer with a sigmoid activation function while the G is encoded by a Deep Convolutional Neural Network (DCNN) with two consecutive fully connected layers. As G is equipped with an encoder-decoder Convolutional Neural Network (CNN), the mode collapse problem of GANs, which refers to not generating varied data, is solved. They used the stochastic gradient descent (SGD) optimisation function with batch normalisation and label smoothing techniques. This model has also been used by Yang et al. [50].

6.1.2. Medical GAN (medGAN)

medGAN was first proposed by Choi et al. [36] as an attempt to solve the mode collapse problem when generating medical binary and categorical data. In this model, an AE is combined with a GAN model to handle binary variables. It is also combined with a minibatch averaging method, batch normalisation and a shortcut connection to the G. Although this method is only compatible with binary data, other authors have used this model to compare or to try to improve its performance. For instance, Jackson and Lussetti [45] proposed an extended version of medGAN (Extended medGAN) to also capture the demographic data of patients. Baowaly et al. [51], Yale et al. [21], Yoon et al. [55] and Lee et al. [56] used medGAN to compare the performance with the method they propose in their studies. In these studies, medGAN gave the poorest results.

In another attempt to improve medGAN, Baowaly et al. [51] created medical Boundary-seeking GAN (medBGAN), a model aimed at producing more accurate discrete and continuous variables. The result was a model with a more stable training and better convergence. Both D and G were composed of feed-forward neural networks.

6.1.3. Deep Convolutional GAN (DCGAN)

DCGAN is an architecture proposed by Park et al. in 2018 [37]. D is a CNN with different layers in which a list of 3x3 learning filters is applied. The last layer has a sigmoid activation function whereas in the other layers batch normalisation is applied and LeakyRelu activation functions are used. G is a neural network (NN) composed of de-convolutional layers and a loss function defined as 'information loss' by the authors. Finally, a classifier, with the same architecture as the D was used to ensure that there is no semantic inconsistency in the generated records. It is compatible with numerical, categorical and time-series data.

6.1.4. Supervised GAN

Norgaard et al. [39] proposed a Supervised GAN model to create synthetic sensor-data. In this model, the G learns from the feedback of both a D and a classifier. The only detail given about the model is that the 3 NNs are composed of a Long Short-Term Memory (LSTM) and a feedforward layer. The D uses a sigmoid activation function while the classifier uses a softmax activation function. In this work, the proposed technique was used to generate time-series sensor data of daily and sports activities.

6.1.5. Sequentially Coupled GAN (SC-GAN)

SC-GAN was proposed by Wang et al. [44] to generate information regarding the status and medication dosage of patients simultaneously. Using coupled generators that interact with each other, the model captures the interactions between the continuous status of patients and the medication dosage they take. It is composed of three components: D (2-layer bidirectional LSTM), patient-status G (2-layer LSTM) and medical dosage G (2-layer LSTM). Medical dosage G generates data with the input of sequential continuous

status of the patients and a random noise vector. The patient-status generator uses previous states, medication dosage and random noise as input. Before this process, both generators are pre-trained to generate the sample of the next time step.

6.1.6. Auxiliary Classifier GAN (AC-GAN)

AC-GAN is a traditional GAN that generates data categorised in the corresponding class. It was proposed by Beaulieu-Jone in 2019 [46]. In this model the G is trained to generate a specified treatment class (standard or intensive) with a random noise vector as input. The generated synthetic records are used to feed the D. The difference with a traditional GAN is that with this model the G knows what type of treatment (class) it should create, as it is specified in the noise vector. These authors also introduce differential privacy (DP) to this method to improve the patient's privacy.

6.1.7. Wasserstein GAN (WGAN)

Wasserstein GAN (WGAN) is an alternative GAN that uses the Wasserstein distance loss function. This loss function is used to solve the mode collapse problem in GANs. This model has been used in many publications, in some cases mixed with DP or other models. For instance, Chin-Cheong et al. [49] used it to generate numerical and categorical data and combined it with DP. Yang et al. [50] used WGAN to compare other GAN-based methods, including their proposal. Baowaly et al. [51], Rashidian et al. [54] and Yoon et al. [55] also used this model to compare with other ones. The last two also added gradient penalty (GP) to the WGAN. Yale et al. [21, 52, 65] also proposed a model based on WGAN and GP in his thesis, named healthGAN, to solve the compatibility and divergence problems of medGAN. SmoothGAN, a model proposed by Rashidian et al. [54], is similarly based on WGAN and GP.

Yoon et al. [55] also used WGAN in an attempt to generate realistic synthetic health care data (ADS-GAN). In this approach, the D minimises the Wasserstein distance while the discriminator measures the Wasserstein distance of both the generated and real records based on GP.

6.1.8. Grouped Correlational GAN (GcGAN)

GcGAN was proposed by Yang et al. [50] to generate more realistic EHRs considering the relationships between multiple diseases, treatments and efficacy. For that, the model is able to group the variables to different categories and explicitly consider their correlations. Composed of one encoder and decoder for each variable group, the G learns the hidden relationships and then uses the decoders to synthesise the data. A batch normalisation layer was used in the G and spectral normalisation in the D. Categorical data was generated in this publication.

6.1.9. Dual Adversarial Autoencoder (DAAE)

DAAE was proposed by Lee et al. [56]. This model combines both an AE and a GAN to learn the continuous latent distribution of the data and the discrete data distribution. It is composed of 3 blocks: seq2seq AE, Inner GAN and Outer GAN. In the first block, the encoder maps an input sequence from the discrete distribution into a continuous code distribution while the decoder generates a model distribution very similar to the original data distribution. The Inner GAN matches the latent distribution with the code distribution. Finally, the Outer GAN adversarially optimises the decoder to produce synthetic distributions fooling the outer D.

6.2. Evaluation Methods for the GAN-based Approaches

In this section the evaluation methods used for GAN-based approaches for resemblance, utility and privacy throughout the reviewed publications are explained to then provide an evaluation and comparison of the GAN-based approaches. From a total of 16 publications in which GAN-based techniques are used, all of them evaluate resemblance, 15 evaluate utility and only 9 evaluate privacy.

6.2.1. Resemblance evaluation

To evaluate the resemblance dimension of SD the following methods and metrics have been used:

- *Compare variable distributions* (Che et al. [35] and Chin-Cheong et al. [49]). If the SD distribution is very similar to RD the resemblance has been classified as "Excellent". Otherwise, if it is more or less similar or not similar, "Good" and "Poor" identifiers have been applied, respectively.
- *Compare frequency of data features* (Che et al. [35] used for the most common 100 features and Yang et al. [50] compared all features). As with the previous one, the categorisation has been completed according to the level of similarity to the RD.
- *Compare dimensional probability or probability distributions* (Choi et al. [36], Wang et al. [44], Abay et al. [48], Baowaly et al. [51] and Yale et al. [21, 52]). With this method, the authors computed the probability distributions of RD and SD attributes and then, compared them. The classification of this metric has been done by analysing the performance of each GAN model to create variables probability more similar to the RD. "Excellent" indicates that the probabilities have been correctly simulated, "Good" indicates that only a few probabilities are maintained in the SD and "Poor" indicates SD does not have the same probabilities as RD.
- *Consultation with clinical experts* (Choi et al. [36], Wang et al. [44], Beaulieu-Jones et al. [46] and Lee et al. [56]) With this method, the authors asked clinical experts to evaluate the SD qualitatively, giving a score between 1 and 10. If the mean score from clinicians is between 1 and 5, the SD resemblance is "Poor", if it is between 5 and 8, resemblance is "Good" and if it is between 8 and 10 resemblance is "Excellent".
- *Data similarity; Synthetic to Synthetic (STS), Real to Synthetic (RTS) and Real to Real (RTR)* Norgaard et al. [39] compared the similarity between segments of data. As the authors do not compare the results obtained with different GAN models and all the average values of STS, RTS and RTR are very similar, the model they proposed has been categorised as "Excellent".
- *Pairwise Pearson correlation coefficient* (Wang et al. [44] and Rashidian et al. [54]). This metric indicates the correlation level between variables. "Excellent" indicates that all the correlations are equal or very similar in RD and SD, "Good" indicates that a few correlations are preserved in SD and "Poor" means that the correlations are not equal in RD and SD.
- *Compare most common values* Jackson and Lussetti [45] used this method to evaluate the resemblance of the unique method they use. As the most common values are equal in RD and generated SD, the approach they proposed has been classified as "Excellent".
- *Compare the mean and standard deviation* (Yang et al. [50] and Rashidian et al. [54]) They calculated the mean and standard deviation Mean absolute error (MAE) between RD and SD. The lower values have been classified as "Excellent", the medium values as "Good" and the higher values as "Poor".
- *Statistical tests* (Baowaly et al. [51] used KS test and Dash et al. [53] used Welsch t-test and Yoon et al. [55] used Student t-test and Chi-squared). In this method the obtained p-value has been analysed and compared to categorise the resemblance of the compared GAN models.
- *Visualisation techniques*: principal component analysis (PCA), histograms and correlation matrices have been used by Yale et al. [21]. Those plots have been visually compared to categorise a model's resemblance as "Excellent" (SD close to RD), "Good" (SD not very close to RD) and "Poor" (SD not close to RD).

- *Nearest neighbor adversarial accuracy (AA) and resemblance loss*. These metrics were proposed by Yale et al. [21, 52]). Nearest neighbor AA indicates the distance between RD and SD, and by combining them resemblance loss can be calculated. In an ideal scenario resemblance loss should be lower than 0.5 to indicate that SD resembles RD. So, a value between 0 and 0.4 indicates "Excellent" resemblance, a value between 0.4 and 0.8 indicates "Good" and higher than 0.8 has been classified as "Poor".
- *Compare average trends*. Dash et al. [53] compared graphically the average trends of SD and RD, indicating that with the proposed model the trends are better preserved in SD ("Excellent" resemblance) than with the SD generated by the other model they compare ("Good" resemblance).
- *Train an ML classifier to label data as real or synthetic*. Lee et al. [56] used this method. In this case, as RD and SD are mixed, the better the model performance, the worse the SD resemblance. So, if an accuracy lower than 0.5 is obtained the resemblance is "Excellent" and if a higher accuracy is obtained the resemblance is "Poor".

6.2.2. Utility evaluation

The utility dimension of SD has been performed with the following methods:

- *Augment data for ML model training* (used by Che et al. [35], Wang et al. [44], Yang et al. [50]). In this case, if the performance of the trained ML model (with a combination of real and synthetic data) has a slight difference in model performance it is considered "Excellent", if the difference is more notable "Good" and if the difference is very big "Poor".
- *Use SD in ML models* The authors use this method in different ways. Both Train on Real Test on Real (TRTR) and Train on Synthetic Test on Real (TSTR) were used by Park et al. [37], Wang et al. [44], Beaulieu-Jones et al. [46], Chin-Cheong et al. [49], Baowaly et al. [51] and Rashidian et al. [54]. Only TSTR was used by Norgaard et al. [39], Abay et al. [48], Yoon et al. [55] and Lee et al. [56]. Dash et al. [53] used all combinations; TRTR, Train on Real Test on Synthetic (TRTS), Train on Synthetic Test on Synthetic (TSTS) and TSTR. Each study used a different metric (e.g. Accuracy, F1-score, ROC, AUC-ROC) to evaluate the performance of the models. To categorise this method more or less the same process as for the previous method has been used; the difference in ML model performance with SD has been evaluated and compared.

6.2.3. Privacy evaluation

The few authors that evaluate privacy dimension use the following methods:

- *Identity disclosure and attribute disclosure* Choi et al. [36] simulated an attack scenario to validate the disclosure of a complete record (identity) and of some attributes. This process was only applied to the SD generated by the method they propose (medGAN). The results showed that if an attacker knows 1% of a patient's attributes, it would be possible to estimate the unknown attributes and the complete patient with 20% accuracy. Therefore, it has been categorised as "Good" because there is a small risk of data disclosure.
- *Distance to the closest record (DCR)*. DCR is the Euclidean distance between an i-record in SD and an i-record in RD. The closer this value is to 0, the better the privacy preservation. Park et al. [37] used this metric to evaluate the privacy of the SD generated with the proposed and compared models. The model with the higher DCR has been categorised as "Good" and the one with the lowest DCR as "Excellent".

- *Membership attack*. Park et al. [37] and Yale et al. [52] simulated a membership attack to see if an attacker could disclose patient data. If in this scenario the privacy is preserved the privacy performance is "Excellent" and if not it is "Good" (whilst privacy model is introduced in the model).
- *Max-RTS similarity*. According to Norgaard et al. [39] the maximum real to synthetic similarity value indicates if the model has memorised and stored RD and is really generating data and not copying it. As this value was never equal to 1, the privacy of the model they proposed has been classified as "Excellent".
- *Formulation of DP*. Beaulieu-Jones et al. [46] used the original formulation of DP, which measures the maximum displacement of the dataset that can be observed by adding or removing a patient. As they compared the same model in two ways (without DP and with DP), the results showed that with DP the generated patients are less similar ("Excellent") to the generated ones with the model that does not apply DP ("Good").
- *Privacy loss*. Yale et al. [21, 52] proposed privacy loss, which is based on the previously defined nearest neighbor AA. The higher the value of the privacy loss, the better the privacy is preserved. So, the privacy dimension of the models they compare have been categorised as "Excellent" for the model with the higher values and as "Poor" for the rest of the models.
- *Maximum mean discrepancy (MMD)*. Rashidian et al. [54] proposed the MMD metric, which is the distance of the space of probabilities of the attributes. If the MMD for SD is lower than for RD, SD is equal to RD and privacy is not preserved. The authors only used this metric to evaluate the privacy of the proposed method, so the requirement to maintain privacy is fulfilled, and the method has been categorised as "Excellent".
- *JS divergence (JSD) and Wasserstein distance*. Yoon et al. [55] computed the balance between identifiability and quality of SD measures in terms of JSD and Wasserstein distance. Privacy has been categorised as "Excellent" if the identifiability is reduced by at least 50% compared with the other models. In the other models, privacy has been considered as "Good" because they use mechanisms to preserve privacy.
- *Differential privacy cost*. Lee et al. [56] analysed the privacy of the proposed method (DAAE) under differential privacy cost induced in the model; a smaller cost value means stronger privacy protection. As they only evaluate privacy in the proposed method and the cost is very small, this model has been categorised as "Excellent".

6.3. Comparison of the GAN-based approaches

In this section the comparison of the GAN-based approaches is exposed in order to answer RQ3. Evaluating and comparing the GAN-based approaches analysed has not been trivial due to the high variability of metrics used in the different works. However, in order to report the results as objectively as possible, the evaluation methods for resemblance, utility and privacy reviewed and described in the previous section are employed to compare the performance of the various techniques. As specified in section 3.2 a per-paper-basis evaluation has been developed to categorise the GAN-based STDG approaches into '*Excellent*", "*Good*" and "*Poor*" in the resemblance, utility and privacy dimensions, based on the results reported for the approaches used in each publication. With this method some conflicts may appear when the same STDG approach has been categorised differently for the same data type in different publications or in a specific dimension. The reason behind this is that in each publication different datasets and different evaluation metrics and methods are used.

Taking the methods and thresholds described in section 6.2 into account, Table 2 shows the results of this comparison. The first column is the publication identifier, the second is the synthesised data type. The third and fourth columns correspond to the size of the data (num. of attributes and num. of records). The fifth column presents the GAN models used in each publication and the last three columns correspond to the performance of each model when evaluating resemblance, utility, and privacy dimensions. If a dimension is not evaluated it is represented by "-". * indicates the winning approach in each publication.

Table 2: Comparative table of publications that used a GAN based approach.

Publication	**Data type**	**Num. of attributes**	**Num. of records**	**GAN based methods**	**Resemblance**	**Utility**	**Privacy**
Che 2017 [35]	Categorical Time-series	-	3357 6714 6744	ehrGAN*	Excellent	Excellent	-
Choi 2017 [36]	Numerical Binary	-	615 1071 569	GAN GANp GANpd GANpa medGAN*	Poor Poor Poor Poor Good	Poor Poor Poor Poor Excellent	- - - - Good
Park 2018 [37]	Numerical Categorical Time-series	23 14 32 32	15000 32561 9813 1000000	table-GAN* DCGAN	Excellent Good	Excellent Poor	Excellent Good
Norgaard 2018 [39]	Numerical Categorical Time-series	480 num. 1 cat.	900	Supervised GAN*	Excellent	Poor	Excellent
Wang 2019 (1) [44]	Categorical Time-series	35 36	13773 5538	SeqGAN C-RNN-GAN RCGAN SC-GAN*	Poor Good Good Excellent	Poor Good Poor Excellent	- - - -
Jackson and Lussetti 2019 [45]	Categorical Time-series	-	10000	Extended medGAN*	Excellent	-	-
Beaulieu-Jones 2019 [46]	Numerical Categorical Time-series	36 45	9361 8260	AC-GAN AC-GAN + DP*	Excellent Excellent	Excellent Excellent	Good Excellent
Chin-Cheong 2019 [49]	Numerical Categorical	795	2873466	WGAN* DP-WGAN	Excellent Good	Excellent Good	Poor Good
Yang 2019 [50]	Categorical	-	17000	WGAN T-WGAN medGAN ehrGAN CorrGAN GcGAN*	Poor Poor Poor Good Good Excellent	Good Good Excellent Excellent Good Excellent	- - - - - -
Baowaly 2019 [51]	Numerical Binary Time-series	942 1651 1015	46517 42214 498909	medGAN medWGAN medBGAN*	Poor Good Excellent	Poor Good Excellent	- - -
Yale 2019 (1) [21]	Discrete Categorical	342	27000	medGAN healthGAN*	Poor Excellent	Poor Excellent	Poor Excellent
Yale 2019 (2) [52]	Discrete Categorical	-	-	healthGAN*	Excellent	Excellent	Excellent
Dash 2020 [53]	Numerical Categorical Binary Time-series	26 34 741	- - -	timeGAN healthGAN*	Good Excellent	- Good	- -

Table 2: Continued on next page

Table 2 (cont)

Publication	Data type	Num. of attributes	Num. of records	GAN based methods	Resemblance	Utility	Privacy
Rashidian 2020 [54]	Numerical Categorical Binary	166	47412	cGAN AC-GAN WGAN WGAN-GP SmoothGAN*	Poor Poor Good Excellent Excellent	Poor Poor Good Excellent Excellent	- - - - Excellent
Yoon 2020 [55]	Numerical Categorical Binary	29 48 41 23	30389 56822 26854 23706	ADS-GAN* PATE-GAN DP-GAN medGAN WGAN-GP	Excellent Good Poor Good Good	Excellent Good Good Good Good	Excellent Good Good Poor Poor
Lee 2020 [56]	Categorical Time-series	39 30	1999 85845	DAAE* medGAN VAE-GAN	Excellent Poor Poor	Excellent Poor Good	Excellent - -

7. Discussion

In this systematic review the recent advances in STDG approaches for PPDP in the healthcare context have been analysed and reported, focusing on the contribution of GAN-based approaches. The outcome of the review has shown many attempts to generate synthetic tabular data to enable secure data exchange in the healthcare context without risk of privacy violation. Thus, the potential of synthetic tabular data to enable progress in the development of AI and knowledge-based decision support healthcare applications (without compromising patient privacy) has been shown. Therefore, the knowledge elicited in this work can be used to gain a detailed insight into current privacy-preserving STDG approaches in the healthcare context. On the one hand, this may be useful for AI researchers looking for a suitable STDG approach to overcome privacy issues in their healthcare application. On the other hand, it may also help guide STDG researchers in identifying and targeting their area of contribution. In the following section the main findings, limitations and research directions emerging from this systematic review are highlighted.

7.1. Main findings

Defining a search strategy to carry out a systematic review in a reproducible way, this paper has answered three RQ (defined in Section 1.2) related to STDG for healthcare tabular data. Next, the findings related with each RQ are presented.

7.1.1. RQ1: What approaches exist for generating synthetic tabular data in healthcare? How can these methods be classified?

From the publications analysed it can be concluded that there is wide variation in the STDG approaches currently used in healthcare , with each approach being appropriate for differing applications and specific objectives. Three main categories of STDG approaches have been identified: classical approaches (baseline methods, statistical models and ML models), DL approaches (AE, GANs and Ensembles) and others. A category has been assigned to each approach by analysing the characteristics and nature of each of them. Based on this categorisation, a complete taxonomy of the most recent STDG approaches has been proposed in Figure 2. This classification highlights the varying characteristics and skills of the identified techniques, suggesting that the appropriate choice of algorithm strongly depends on the application and the objective of STDG.

Classical approaches have been found to be weak in generating high quality synthetic tabular data that protects the privacy of real data. Usability and privacy of the data generated with these approaches is not well balanced as they typically try to memorise real data and the relationships between attributes . However, they have been widely used as a baseline to compare newer and more complex technologies. DL approaches have shown promising results in generating higher quality and better privacy-preserving synthetic tabular data as they perform better in learning patterns of real data and in generating more diverse data. For this reason, the popularity of DL-based, and especially GAN-based approaches', has significantly increased in recent years for privacy-preserving STDG. The methods included in the last group (i.e., others) have shown good results for the applications they were designed for, however, their suitability for other applications and contexts is yet to be analysed, because they use predefined data schemes designed for the application of interest."

7.1.2. RQ2: Which of these approaches are based on GANs? What are their characteristics and/or distinctions?

In total 23 GAN-based approaches have been found for STDG from the retrieved publications. The majority of these works are existing solutions' based improved approaches through the addition or modification of certain characteristics to adapt them to the data type of interest. To give a clearer answer to this RQ, the approaches have been clustered into 9 groups. The approaches in the different groups differ mainly in the architecture of the generator and discriminator, as well as on their parameters. The GAN-based approaches that are more suitable for time-series data are LSTM networks, while the other approaches are usually composed of CNNs. Another distinctive feature among these approaches is that in some cases the Wasserstein distance is used to generate more diverse data (to solve the mode collapse problem of GANs). The conclusion that can be drawn from this finding is that the architectures, parameters and additions that can be applied and evaluated for a GAN-based approach are so varied that it takes a long time to get an ideal approach for the data of interest and that the existing GAN-based approaches are not generalisable to all types of tabular data.

7.1.3. RQ3: What is the performance of these approaches in creating data that are usable, private and similar but not identical to real data?

The approaches that are based on GANs have been analysed and evaluated, considering the data type they use and the performance of those models in the resemblance, utility and privacy dimensions. The interpretation of dimension metrics and their evaluation has been done on an article-by-article basis, as the metrics used for the evaluation of synthetic tabular data differ among publications.. Therefore, it is not possible to get a clear view of the best GAN-based approach and conflicts may arise for the same data type in different publications or dimensions. However, from Table 2 it can be concluded that, in general, GAN-based approaches perform very well when generating data that are usable, private and similar but not identical to real data. It has also been found that in most publications, with the evaluation methods used, the proposed GAN-based approach outperforms the other approaches used for comparison.

The research has also shown that the privacy dimension is usually not evaluated. Furthermore, the performance dimension (i.e., footprint and computational cost) is only evaluated in a small number of publications, so there is not a clear metric or method to evaluate it. Finally, one of the most significant findings to emerge from this study is that there are no standardised metrics or methods to evaluate and benchmark the different approaches for resemblance, utility and privacy dimensions."

7.2. Limitations

A limitation of this study is that the in-depth analysis of the resemblance, utility and privacy characteristics was only performed for GAN-based techniques. Given the existence of such a wide variety of STDG ap-

proaches it has been necessary to focus on a subset, and GANs have been identified as the best candidates in this regard due to their performance and thus popularity in other areas and applications. Moreover, the major limitation of the study is the lack of comparison of identified GAN models due to the absence of standard objective metrics for the three analysed dimensions (resemblance, privacy and utility). To overcome this issue, the current paper has proposed an alternative categorisation methodology and has established some criterion to evaluate the "Poor", "Good" or "Excellent" performance in each of the analysed dimensions for each publication. As in each publication different datasets and different evaluation metrics and methods are used, this method may be conflicting when the same STDG approach has been categorised differently for the same data type in a specific dimension. An additional uncontrolled factor is the possibility that the results reported in each study included in the systematic review, could have been biased and/or favourable to the proposed approach by strategically selecting the metrics to show. Notwithstanding these limitations, the study suggests a taxonomy of the current STDG algorithms, focused on GANs, while giving a comprehensive review of their characteristics that can be very useful and can help researchers select the appropriate technique for their applications or use it as a starting point for replicable metrics development.

7.3. Research directions

This research has raised many questions in need of further investigation. First, most of the reviewed techniques lack assessments regarding the three analysed dimensions and the performance dimension. Therefore, further research should focus on determining the performance of the existing techniques in these dimensions. Towards that goal, it is essential to define standard, objective and reliable metrics and benchmarks to evaluate the selected skills (resemblance, utility, privacy and performance). Further studies where the different algorithms face the same problem, starting from a strategic dataset that requires a wide range of capabilities (generation of categorical, numerical, temporal data etc.), can help in reliably establishing the benchmark and making the assessments against it. The definition of such a benchmark will also highlight the need for further developments and improvements in STDG for tabular applications, in order to reach or improve the numbers already achieved in other fields.

Second, it has been found that there is not a best GAN-based approach to generate synthetic tabular data since each GAN model has its strengths and weaknesses and is adapted to the data type of interest. Therefore, further work on GAN models should be performed to improve the generalisability of GANs to find a model that works optimally across all types of tabular healthcare data.

8. Conclusion

In the era of digitalisation and Big Data, the development and application of innovative AI algorithms is the order of the day. However, for this technology to be developed and used in applications from which the entire society can benefit, it is essential that the scientific community is given access to the data. Unfortunately, problems related to privacy, intellectual property and security place many barriers on data sharing. This is magnified in the health sector, where data are often particularly sensitive. Techniques that allow data sharing in preserving the security and privacy of individuals, and does not violate any of their rights, are a priority so that research in this area is not affected. STDG techniques have shown promise in providing an effective solution to this problem.

The main goal of the current systematic review was to give an overview of the STDG approaches that could be useful in healthcare applications where tabular data needs to be generated, giving special attention to GANs due to their recent success for similar purposes. This study has found that three main groups can be distinguished when classifying the reviewed STDG algorithms, but that many of them lack an evaluation

of their privacy-preserving skills and performance dimensions. This is something essential along with the definition of objective measurement standards and metrics that do not yet exist.

The findings from this study make several contributions to the current literature. First, this work has been one of the first attempts to thoroughly examine and classify the existing STDG techniques for tabular data with applications in the healthcare domain, focusing especially on GANs and evaluating their resemblance, utility and privacy dimensions. The insights gained from this study may be of assistance to the scientific community, especially to those researchers who are willing to develop AI-based applications without compromising users' privacy or who have difficulties in advancing their research fields due to these issues.

It is unfortunate that the study may have been affected by the lack of standardisation in the metrics reported in the literature, a limitation that has been overcome with a new proposal for comparisons. Further work needs to be done to define standard benchmarks and metrics for the evaluation of STDG algorithms' dimensions, as well as to keep developing and improving the algorithms themselves. Only the widespread use of fully synthetic tabular data will enable progress in the development of intelligent health applications, always preserving the privacy of the patients.

Appendix A. Abbreviations

AA	Adversarial Accuracy
AC-GAN	Auxiliary Classifier Generative Adversarial Network
actGAN	Activation-Specific Generative Adversarial Network
ADS-GAN	Anonymization through Data Synthesis using Generative Adversarial Networks
AE	Autoencoder
AI	Artificial Intelligence
ANM	Additional Noise Model
ARAE	Adversarially Regularized Auto-Encoder
ASD	Autism Spectrum Disorder
BN	Bayesian Network
CART	Classification and Regression Trees
CLGP	Categorical Latent Gaussian Process
CMC	Contraceptive Method Choice
CMEM	Categorical Maximum Entropy Model
CoMSER	Content Modelling for Synthetic E-Health Records
CorrGAN	Correlational Generative Adversarial Network
CRD	Copy Real Data
CVD	Cardiovascular Disease
DBM	Deep Boltzmann Machines
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCGAN	Deep Convolutional Generative Adversarial Network
DCR	Distance to the Closest Record
DP	Differential Privacy
DTW	Dynamic Time Wrapping
EHR	Electronic Health Records
GAN	Generative Adversarial Network
GcGAN	Grouped Correlational Generative Adversarial Network
ehrGAN	Electronic health records Generative Adversarial Network

EM	Expectation Maximization
GM	Gaussian Multivariate
ICD9	International Classification of Diseases, 9th version
ICU	Intensive Care Unit
IS	Independent Sampling
JSD	Jensen-Shannon Divergence
JSON	JavaScript Object Notation
KDE	Kernel Density Estimation
KS	Kolmogorov-Smirnov
LOG. REG.	Logistic Regression
LR	Linear Regression
MAE	Mean Absolute Error
MAGGIC	Meta-Analysis Global Group in Chronic Heart Failure
MC-medGAN	Multicategorical Medical Generative Adversarial Network
MICE-DT	
MKDE	Multivariate Kernel Density Estimation
medBGAN	Medical Boundary-seeking Generative Adversarial Network
medGAN	Medical Generative Adversarial Network
medWGAN	Medical Wasserstein Generative Adversarial Network
ML	Machine Learning
MMD	Mean Maximum Discrepancy
NHIRD	National Health Insurance Research
NMDS	Non-metric Multidimensional Scaling
MPoM	Mixture of product of Multinomials
PATE-GAN	Private Aggregation of Teacher Ensembles Generative Adversarial Network
PCA	Principal Component Analysis
PCD	Pairwise Correlation Difference
PPDP	Privacy-Preserving Data Publishing
PW	Parzen Window
PrivBayes	Private Naive Bayes
RCGAN	Recurrent Convolutional Generative Adversarial Network
RD	Real Data
RD	Random Decision Trees
RN	Random Noise
RQ	Research Questions
RTR	Real To Real
RTS	Real To Synthetic
SC-GAN	Sequentially Coupled Generative Adversarial Network
SeqGAN	Sequential Generative Adversarial Network
SD	Synthetic Data
SDG	Synthetic Data Generation
SPRINT	Systolic Blood Pressure Intervention Trial
STDG	Synthetic Tabular Data Generation
STS	Synthetic To Synthetic
SVM	Support Vector Machines
TRTR	Train on Real and Test on Real

TRTS	Train on Real and Test on Synthetic
TSTR	Train on Synthetic and Test on Real
TSTS	Train on Synthetic and Test on Synthetic
UNOS	United Network for Organ Transplantation
VAE	Variational Auto-Encoder
WGAN	Wasserstein Generative Adversarial Network
WIE	Weighted Itemset Error

References

[1] G. Epelde, A. Beristain, R. Alvarez, M. Arrúe, I. Ezkerra, O. Belar, R. Bilbao, G. Nikolic, X. Shi, B. D. Moor, M. Mulvenna, Quality of data measurements in the big data era: Lessons learned from MIDAS project, IEEE Instrumentation Measurement Magazine 23 (2020) 18–24. doi:10.1109/MIM.2020.9234761.

[2] C. C. Aggarwal, P. S. Yu, A general survey of privacy-preserving data mining models and algorithms, in: C. C. Aggarwal, P. S. Yu (Eds.), Privacy-Preserving Data Mining: Models and Algorithms, Springer US, 2008, pp. 11–52. doi:10.1007/978-0-387-70992-5_2.

[3] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology 10 (2019) 12:1–12:19. doi:10.1145/3298981.

[4] A. Alabdulatif, I. Khalil, X. Yi, Towards secure big data analytic for cloud-enabled applications with fully homomorphic encryption, Journal of Parallel and Distributed Computing 137 (2020) 192–204. doi:10.1016/j.jpdc.2019.10.008.

[5] B. C. M. Fung, K. Wang, R. Chen, P. S. Yu, Privacy-preserving data publishing: A survey of recent developments, ACM Computing Surveys 42 (2010) 14:1–14:53. doi:10.1145/1749603.1749605.

[6] K. Singh, L. Batten, Aggregating privatized medical data for secure querying applications, Future Generation Computer Systems 72 (2017) 250–263. doi:10.1016/j.future.2016.11.028.

[7] P. Li, T. Li, H. Ye, J. Li, X. Chen, Y. Xiang, Privacy-preserving machine learning with multiple data providers, Future Generation Computer Systems 87 (2018) 341–350. doi:10.1016/j.future.2018.04.076.

[8] J. P. Reiter, New approaches to data dissemination: A glimpse into the future (?), CHANCE 17 (2004) 11–15. doi:10.1080/09332480.2004.10554907, publisher: Taylor & Francis.

[9] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, G. Epelde, Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing, JMIR Medical Informatics 8 (2020) e18910. doi:10.2196/18910.

[10] D. Migdal, C. Rosenberger, Statistical modeling of keystroke dynamics samples for the generation of synthetic datasets, Future Generation Computer Systems 100 (2019) 907–920. doi:10.1016/j.future.2019.03.056.

[11] X. Liu, N. Iftikhar, H. Huo, R. Li, P. S. Nielsen, Two approaches for synthesizing scalable residential energy consumption data, Future Generation Computer Systems 95 (2019) 586–600. doi:10.1016/j.future.2019.01.045.

[12] T. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, Future Healthcare Journal 6 (2019) 94–98. doi:10.7861/futurehosp.6-2-94.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 27, Curran Associates, Inc., 2014, pp. 2672–2680.

[14] H. Alqahtani, M. Kavakli-Thorne, G. Kumar, Applications of generative adversarial networks (GANs): An updated review, Archives of Computational Methods in Engineering (2019). doi:10.1007/s11831-019-09388-y.

[15] J. Georges-Filteau, E. Cirillo, Synthetic observational health data with GANs: from slow adoption to a boom in medical research and ultimately digital twins?, Authorea, Inc. (2020). doi:10.21203/rs.3.rs-116297/v2.

[16] K. El Emam, R. Hoptroff, The Synthetic Data Paradigm for Using and Sharing Data, DATA ANALYTICS & DIGITAL TECHNOLOGIES 19 (2019) 12.

[17] A. Hernandez-Matamoros, H. Fujita, H. Perez-Meana, A novel approach to create synthetic biomedical signals using BiRNN, Information Sciences 541 (2020) 218–241. doi:10.1016/j.ins.2020.06.019.

[18] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, H. Nakayama, GAN-based synthetic brain MR image generation, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 734–738. doi:10.1109/ISBI.2018.8363678, ISSN: 1945-8452.

[19] J. Guan, R. Li, S. Yu, X. Zhang, Generation of synthetic electronic medical record text, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 374–380. doi:10.1109/BIBM.2018.8621223.

[20] J. Dahmen, D. Cook, SynSys: A synthetic data generation system for healthcare applications, Sensors 19 (2019) 1181. doi:10.3390/s19051181, number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

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

[22] D. Zhang, C. Yin, J. Zeng, X. Yuan, P. Zhang, Combining structured and unstructured data for predictive models: a deep learning approach, BMC Medical Informatics and Decision Making 20 (2020) 280. doi:10.1186/s12911-020-01297-6.

[23] K. M. Chong, Privacy-preserving healthcare informatics: a review, ITM Web of Conferences 36 (2021) 04005. doi:10.1051/itmconf/20213604005, publisher: EDP Sciences.

[24] M. S. Donaldson, K. N. Lohr, I. o. M. U. Committee on Regional Health Data, Health Data in the Information Age: Use, Disclosure, and Privacy, National Academies Press (US), 1994, pp. 142–144.

[25] W. N. Price, I. G. Cohen, Privacy in the age of medical big data, Nature Medicine 25 (2019) 37–43. doi:10.1038/s41591-018-0272-7.

[26] H.-Y. Tran, J. Hu, Privacy-preserving big data analytics a comprehensive survey, Journal of Parallel and Distributed Computing 134 (2019) 207–218. URL: https://www.sciencedirect.com/science/article/pii/S0743731519300589. doi:10.1016/j.jpdc.2019.08.007.

[27] S. Wang, L. Bonomi, W. Dai, F. Chen, C. Cheung, C. S. Bloss, S. Cheng, X. Jiang, Big data privacy in biomedical research, IEEE Transactions on Big Data 6 (2020) 296–308. doi:10.1109/TBDATA.2016.2608848, conference Name: IEEE Transactions on Big Data.

[28] A. Chester, Y. S. Koh, J. Wicker, Q. Sun, J. Lee, Balancing utility and fairness against privacy in medical data, in: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2020, pp. 1226–1233. doi:10.1109/SSCI47803.2020.9308226.

[29] Z. Azizi, C. Zheng, L. Mosquera, L. Pilote, K. E. Emam, Can synthetic data be a proxy for real clinical trial data? a validation study, BMJ Open 11 (2021) e043497. doi:10.1136/bmjopen-2020-043497, publisher: British Medical Journal Publishing Group Section: Health informatics.

[30] K. E. Emam, L. Mosquera, J. Bass, Evaluating identity disclosure risk in fully synthetic health data: Model development and validation, Journal of Medical Internet Research 22 (2020) e23139. doi:10.2196/23139, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[31] F. K. Dankar, M. Ibrahim, Fake it till you make it: Guidelines for effective synthetic data generation, Applied Sciences 11 (2021) 2158. URL: https://www.mdpi.com/2076-3417/11/5/2158. doi:10.3390/app11052158, number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[32] K. S. Khan, R. Kunz, J. Kleijnen, G. Antes, Five steps to conducting a systematic review, Journal of the Royal Society of Medicine 96 (2003) 4.

[33] L. S. Uman, Systematic reviews and meta-analyses, Information Management for the Busy Practitioner (2011) 3.

[34] S. McLachlan, K. Dube, T. Gallagher, Using the CareMap with health incidents statistics for generating the realistic synthetic electronic healthcare record, in: 2016 IEEE International Conference on Healthcare Informatics (ICHI), 2016, pp. 439–448. doi:10.1109/ICHI.2016.83.

[35] Z. Che, Y. Cheng, S. Zhai, Z. Sun, Y. Liu, Boosting deep learning risk prediction with generative adversarial networks for electronic health records, in: 2017 IEEE International Conference on Data Mining (ICDM), 2017, pp. 787–792. doi:10.1109/ICDM.2017.93, ISSN: 2374-8486.

[36] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, J. Sun, Generating multi-label discrete patient records using generative adversarial networks, Proceedings of Machine Learning for Healthcare (MLHC) (2018).

[37] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, Y. Kim, Data synthesis based on generative adversarial networks, Proceedings of the VLDB Endowment 11 (2018) 1071–1083. doi:10.14778/3231751.3231757.

[38] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, S. 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 (2018) 230–238. doi:10.1093/jamia/ocx079.

[39] S. Norgaard, R. Saeedi, K. Sasani, A. H. Gebremedhin, Synthetic sensor data generation for health applications: A supervised deep learning approach, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 1164–1167. doi:10.1109/EMBC.2018.8512470, ISSN: 1558-4615.

[40] H. Wu, Y. Ning, P. Chakraborty, J. Vreeken, N. Tatti, N. Ramakrishnan, Generating realistic synthetic population datasets, ACM Transactions on Knowledge Discovery from Data 12 (2018) 45:1–45:22. doi:10.1145/3182383.

[41] M. Zare, J. Wojtusiak, Weighted itemsets error (WIE) approach for evaluating generated synthetic patient data, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 1017–1022. doi:10.1109/ICMLA.2018.00166.

[42] J. Vaidya, B. Shafiq, M. Asani, N. Adam, X. Jiang, L. Ohno-Machado, A scalable privacy-preserving data generation methodology for exploratory analysis, AMIA Annual Symposium Proceedings 2017 (2018) 1695–1704.

[43] S. McLachlan, K. Dube, T. Gallagher, J. A. Simmonds, N. Fenton, Realistic synthetic data generation: The ATEN framework, in: A. Cliquet Jr., S. Wiebe, P. Anderson, G. Saggio, R. Zwiggelaar, H. Gamboa, A. Fred, S. Bermúdez i Badia (Eds.), Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science, Springer International Publishing, 2019, pp. 497–523. doi:10.1007/978-3-030-29196-9_25.

[44] L. Wang, W. Zhang, X. He, Continuous patient-centric sequence generation via sequentially coupled adversarial learning, in:

G. Li, J. Yang, J. Gama, J. Natwichai, Y. Tong (Eds.), Database Systems for Advanced Applications, Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 36–52. doi:10.1007/978-3-030-18579-4_3.

[45] P. Jackson, M. Lussetti, Extending a generative adversarial network to produce medical records with demographic characteristics and health system use, in: 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2019, pp. 0515–0518. doi:10.1109/IEMCON.2019.8936168, ISSN: 2644-3163.

[46] Beaulieu-Jones Brett K., Wu Zhiwei Steven, Williams Chris, Lee Ran, Bhavnani Sanjeev P., Byrd James Brian, Greene Casey S., Privacy-preserving generative deep neural networks support clinical data sharing, Circulation: Cardiovascular Quality and Outcomes 12 (2019) e005122. doi:10.1161/CIRCOUTCOMES.118.005122.

[47] Z. Wang, P. Myles, A. Tucker, Generating and evaluating synthetic UK primary care data: Preserving data utility patient privacy, in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), 2019, pp. 126–131. doi:10.1109/CBMS.2019.00036, ISSN: 2372-9198.

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

[49] K. Chin-Cheong, T. Sutter, J. E. Vogt, Generation of heterogeneous synthetic electronic health records using GANs, in: Workshop on Machine Learning for Health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), ETH Zurich, Institute for Machine Learning, 2019, pp. 1–6. doi:10.3929/ethz-b-000392473, accepted: 2020-06-16T05:38:56Z.

[50] F. Yang, Z. Yu, Y. Liang, X. Gan, K. Lin, Q. Zou, Y. Zeng, Grouped correlational generative adversarial networks for discrete electronic health records, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 906–913. doi:10.1109/BIBM47256.2019.8983215.

[51] M. K. Baowaly, C.-C. Lin, C.-L. Liu, K.-T. Chen, Synthesizing electronic health records using improved generative adversarial networks, Journal of the American Medical Informatics Association 26 (2019) 228–241. doi:10.1093/jamia/ocy142.

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

[53] S. Dash, A. Yale, I. Guyon, K. P. Bennett, Medical time-series data generation using generative adversarial networks, in: M. Michalowski, R. Moskovitch (Eds.), Artificial Intelligence in Medicine, Lecture Notes in Computer Science, Springer International Publishing, 2020, pp. 382–391. doi:10.1007/978-3-030-59137-3_34.

[54] S. Rashidian, F. Wang, R. Moffitt, V. Garcia, A. Dutt, W. Chang, V. Pandya, J. Hajagos, M. Saltz, J. Saltz, SMOOTH-GAN: Towards sharp and smooth synthetic EHR data generation, in: M. Michalowski, R. Moskovitch (Eds.), Artificial Intelligence in Medicine, Lecture Notes in Computer Science, Springer International Publishing, 2020, pp. 37–48. doi:10.1007/978-3-030-59137-3_4.

[55] J. Yoon, L. N. Drumright, M. v. d. Schaar, Anonymization through data synthesis using generative adversarial networks (ADS-GAN), IEEE Journal of Biomedical and Health Informatics 24 (2020) 2378–2388. doi:10.1109/JBHI.2020.2980262.

[56] D. Lee, H. Yu, X. Jiang, D. Rogith, M. Gudala, M. Tejani, Q. Zhang, L. Xiong, Generating sequential electronic health records using dual adversarial autoencoder, Journal of the American Medical Informatics Association 27 (2020) 1411–1419. doi:10.1093/jamia/ocaa119, publisher: Oxford Academic.

[57] E. E. Fowler, A. Berglund, M. J. Schell, T. A. Sellers, S. Eschrich, J. Heine, Empirically-derived synthetic populations to mitigate small sample sizes, Journal of Biomedical Informatics 105 (2020) 103408. doi:10.1016/j.jbi.2020.103408.

[58] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, A. P. Sales, Generation and evaluation of synthetic patient data, BMC Medical Research Methodology 20 (2020) 108. URL: https://doi.org/10.1186/s12874-020-00977-1. doi:10.1186/s12874-020-00977-1.

[59] J. Hyun, S. H. Lee, H. M. Son, J.-U. Park, T.-M. Chung, A synthetic data generation model for diabetic foot treatment, in: T. K. Dang, J. Küng, M. Takizawa, T. M. Chung (Eds.), Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications, Communications in Computer and Information Science, Springer, 2020, pp. 249–264. doi:10.1007/978-981-33-4370-2_18.

[60] S. Wang, C. Rudolph, S. Nepal, M. Grobler, S. Chen, PART-GAN: privacy-preserving time-series sharing, in: Artificial Neural Networks and Machine Learning – ICANN 2020: 29th International Conference on Artificial Neural Networks Bratislava, Slovakia, September 15–18, 2020 Proceedings, Part I, Springer, 2020, pp. 578–593. doi:10.1007/978-3-030-61609-0_46.

[61] A. Koivu, M. Sairanen, A. Airola, T. Pahikkala, Synthetic minority oversampling of vital statistics data with generative adversarial networks, Journal of the American Medical Informatics Association 27 (2020) 1667–1674. doi:10.1093/jamia/ocaa127.

[62] A. Tucker, Z. Wang, Y. Rotalinti, P. Myles, Generating high-fidelity synthetic patient data for assessing machine learning healthcare software, npj Digital Medicine 3 (2020) 1–13. URL: https://www.nature.com/articles/s41746-020-00353-9. doi:10.1038/s41746-020-00353-9, number: 1 Publisher: Nature Publishing Group.

[63] Z. Wang, P. Myles, A. Tucker, Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy, Computational Intelligence 37 (2021) 819–851. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12427. doi:https://doi.org/10.1111/coin.12427, _eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/coin.12427.

[64] Z. Zhang, C. Yan, T. A. Lasko, J. Sun, B. A. Malin, SynTEG: a framework for temporal structured electronic health data simulation, Journal of the American Medical Informatics Association 28 (2021) 596–604. URL: https://doi.org/10.1093/jamia/ocaa262. doi:10.1093/jamia/ocaa262.

[65] A. J. Yale, Privacy Preserving Synthetic Health Data Generation and Evaluation, Ph.D. thesis, Rensselaer Polytechnic Institute, 2020. ISBN: 9798662575981 Publication Title: ProQuest Dissertations and Theses 27833340.

[66] J. Chen, D. Chun, M. Patel, E. Chiang, J. James, The validity of synthetic clinical data: a validation study of a leading synthetic data generator (synthea) using clinical quality measures, BMC Medical Informatics and Decision Making 19 (2019) 44. doi:10.1186/s12911-019-0793-0.

[67] J. Walonoski, S. Klaus, E. Granger, D. Hall, A. Gregorowicz, G. Neyarapally, A. Watson, J. Eastman, Synthea™ novel coronavirus (COVID-19) model and synthetic data set, Intelligence-Based Medicine 1-2 (2020) 100007. doi:10.1016/j.ibmed.2020.100007.

Mikel Hernandez

Mikel Hernandez holds a degree in Telecommunication Systems Engineering (2019) and a master's degree in Biomedical Technologies (2021) from the Faculty of Engineering of Mondragon University. Currently, is working as a Research Assistant in the Precision Medicine and Big Data line of the Digital Health and Biomedical Technologies department at the Vicomtech foundation, on which he developed his final master's degree thesis. His work is related with synthetic data generation and evaluation in the healthcare and industry contexts.

Gorka Epelde

Gorka Epelde is a Project Leader and Senior Researcher in Vicomtech. In 2014, Gorka obtained his Computer Science PhD from the University of the Basque Country. From 2000 until 2007 Gorka held the position of Assistant Researcher at Ikerlan. From 2007 onwards, Gorka has been a Staff Researcher at Vicomtech's eHealth and Biomedical Applications department. Since 2009, he is part of the eHealth group under the Bioengineering Area of the BioDonostia Health Research Institute. His fields of interest include interoperability architectures, data engineering, as well as the human computer interaction and the advanced visualisation of data.

Ane Alberdi

Lecturer and researcher at MGEP since 2017. Double graduate in Electronics Engineering from INP-ENSEEIHT and MGEP, she also holds a Master's degree in Embedded Systems. She received her Ph.D. in 2017, where she worked on the early detection of diseases using smart environments (SE) and AI techniques. As a visiting researcher at the CASAS Lab. at Washington State University (WA, USA) during her Ph.D., she studied the use of SE to support the independent living of the cognitively impaired elderly. Ane has published articles in several scientific journals of impact. Currently, her research is focused in health-related applications of AI.

Rodrigo Cilla

Rodrigo Cilla holds a degree in Computer Engineering (2007), a Master in Computer Science and Technology (2008) and a PhD (2012) from Universidad Carlos III de Madrid. He worked as a postdoctoral researcher in Instituto de Telecomunicaçoes (Lisbon, Portugal, 2013), as a Research Fellow at the Victor Hatini Lab of Tufts University (2013 - 2016) and as a postdoctoral researcher at Fetal I+D Research Center of Fundació Sant Joan

de Dèu (Espluges de Llobregat, Catalonia, 2016-2018). Since 2018 he is a researcher at Fundación Vicomtech (Donostia, Basque Country) where he is involved in different projects related to biomedical computer vision and machine learning.

Debbie Rankin

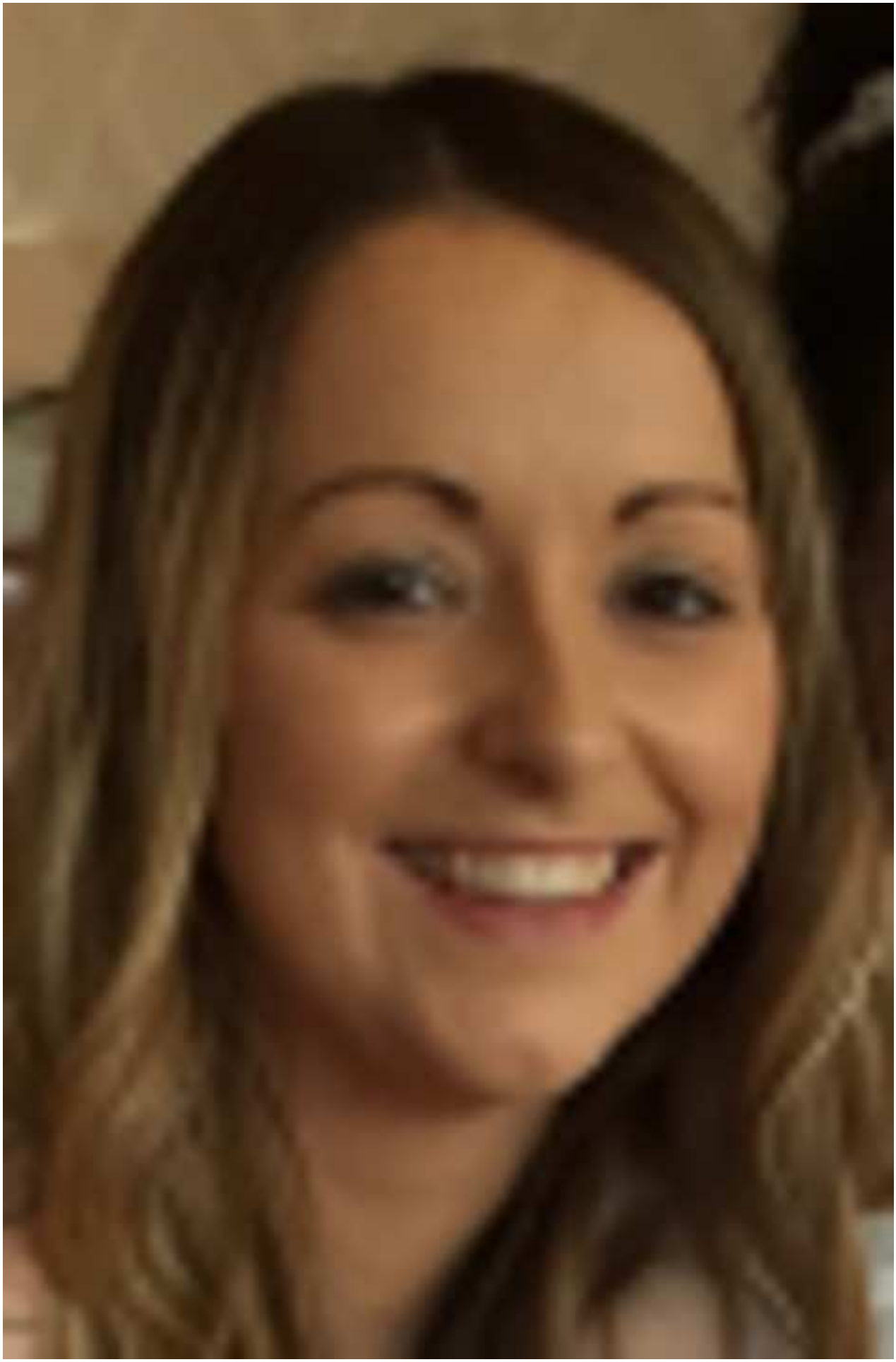
Debbie Rankin received her B.Sc. and Ph.D. degrees in Computer Science from Ulster University, Northern Ireland (UK) in 2008 and 2012, respectively. She is currently a Lecturer in Computer Science at Ulster University in the School of Computing, Engineering and Intelligent Systems. Her research covers Health Informatics, Synthetic Data, Machine Learning, Data Mining, Big Data and Computer Vision.

Ane Alberdi

Click here to access/download;Photo of the author(s);AneAlberdi.jpg

Debbie Rankin

Click here to access/download;Photo of the author(s);DebbieRankin.png



Mikel Hernandez

Click here to access/download;Photo of the author(s);MikelHernandez.JPG

Rodrigo Cilla

Click here to access/download;Photo of the author(s);RodrigoCilla.jpg

Gorka Epelde

Click here to access/download;Photo of the author(s);Gorka Epelde.jpg