Journal of Biomedical Informatics 144 (2023) 104436

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV

Nicholas I-Hsien Kuo [a,*], Federico Garcia [b,c,d], Anders Sönnerborg [e], Michael Böhm [f], Rolf Kaiser [f], Maurizio Zazzi [g], EuResist Network study group, Mark Polizzotto [h], Louisa Jorm [a], Sebastiano Barbieri [a]

[a] Centre for Big Data Research in Health, the University of New South Wales, Sydney, Australia
[b] Instituto de Investigación Ibs.Granada, Spain
[c] Hospital Universitario San Cecilio, Spain
[d] CIBER de Enfermedades Infecciosas, Spain
[e] Hospital Karolinska Institutet, Sweden
[f] Uniklinik Köln, Universität zu Köln, Germany
[g] Università degli Studi di Siena, Italy
[h] Australian National University, Canberra, Australia

ARTICLE INFO

Keywords:
Machine learning
Generative adversarial networks
Human immunodeficiency virus

ABSTRACT

Objective: Clinical data's confidential nature often limits the development of machine learning models in healthcare. Generative adversarial networks (GANs) can synthesise realistic datasets, but suffer from mode collapse, resulting in low diversity and bias towards majority demographics and common clinical practices. This work proposes an extension to the classic GAN framework that includes a variational autoencoder (VAE) and an external memory mechanism to overcome these limitations and generate synthetic data accurately describing imbalanced class distributions commonly found in clinical variables.

Methods: The proposed method generated a synthetic dataset related to antiretroviral therapy for human immunodeficiency virus (ART for HIV). We evaluated it based on five metrics: (1) accurately representing imbalanced class distribution; (2) the realism of the individual variables; (3) the realism among variables; (4) patient disclosure risk; and (5) the utility of the generated dataset for developing downstream machine learning models.

Results: The proposed method overcomes the issue of mode collapse and generates a synthetic dataset that accurately describes imbalanced class distributions commonly found in clinical variables. The generated data has a patient disclosure risk of 0.095%, lower than the 9% threshold stated by Health Canada and the European Medicines Agency, making it suitable for distribution to the research community with high security. The generated data also has high utility, indicating the potential of the proposed method to enable the development of downstream machine learning algorithms for healthcare applications using synthetic data.

Conclusion: Our proposed extension to the classic GAN framework, which includes a VAE and an external memory mechanism, represents a promising approach towards generating synthetic data that accurately describe imbalanced class distributions commonly found in clinical variables. This method overcomes the limitations of GANs and creates more realistic datasets with higher patient cohort diversity, facilitating the development of downstream machine learning algorithms for healthcare applications.

1. Introduction

Healthcare data are indispensable for developing machine learning models that assist clinical decision-making. Yet, acquiring these data is often impeded due to the confidentiality of patient information [1] and regulations safeguarding patient rights [2–4]. This limitation, in turn, curtails the sharing, comparison, and systematic progress in machine learning applications in healthcare.

* Corresponding author.
E-mail address: n.kuo@unsw.edu.au (N.I.-H. Kuo).

https://doi.org/10.1016/j.jbi.2023.104436
Received 31 January 2023; Received in revised form 24 June 2023; Accepted 30 June 2023
Available online 13 July 2023

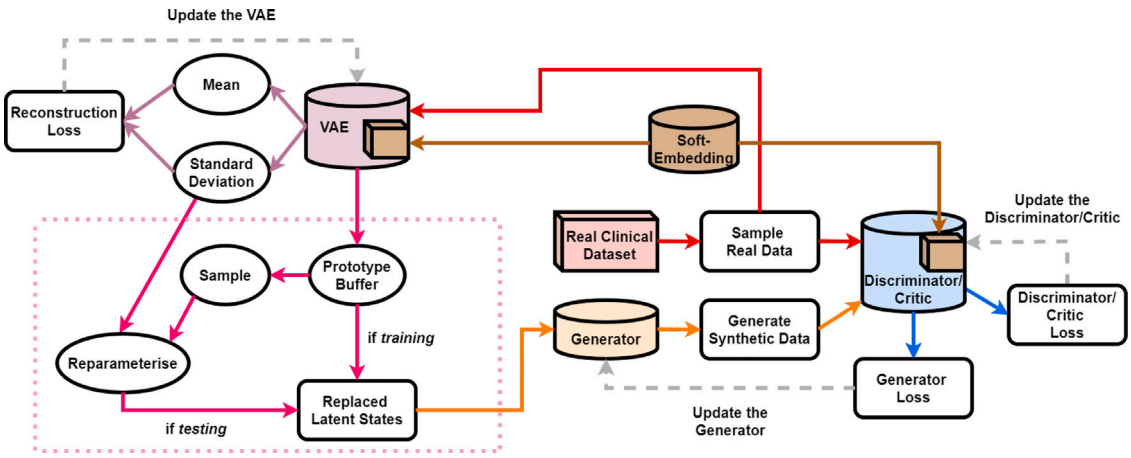


Fig. 1. Extending GAN with a buffer which replays observed real data features.

Procuring prospective clinical data is a time-consuming and expensive undertaking [5]. Synthetic data, as a result, offers a promising alternative for preliminary machine learning model prototyping [6], thus streamlining the collection of real data and offering insights into the optimal type and method of data collection. Besides, synthetic data can attenuate privacy risks inherent in the distribution of real clinical data [7].

A central machine learning paradigm, *reinforcement learning* (RL) [8], is evolving rapidly and holds immense potential for aiding medical practitioners in complex decision-making tasks by offering evidence-based clinical decision support [9,10]. Yet, the extensive datasets required for successful RL algorithm development are seldom available in healthcare. This data scarcity is not confined to RL, it also affects traditional statistical studies and other machine learning disciplines.

To circumvent the insufficient amount of healthcare data, generative machine learning models offer a promising solution by synthesising datasets that can supplant sensitive real data. *Generative adversarial networks* (GANs) [11–13] are a potent tool capable of generating synthetic data mirroring the characteristics of real data. Such GAN-generated datasets can be freely shared, supporting machine learning model benchmarking and healthcare education [14]. However, GANs are prone to *mode collapse* [15], a frequent issue that hampers their effectiveness.

Mode collapse occurs when a GAN produces synthetic data with limited diversity. Consequently, the synthetic dataset may overlook uncommon treatments or, more alarmingly, omit patients from underrepresented backgrounds. This diminishes the generated dataset's utility and may indirectly lead to patient harm [16].

In our paper, we propose an augmented GAN framework designed to tackle mode collapse in the synthesis of real-world clinical datasets. Our novel method, as illustrated in Fig. 1, incorporates a *variational autoencoder* (VAE) [17] and an external buffer to store features encoded from real data. Focusing on antiretroviral therapy for human immunodeficiency virus (ART for HIV), our experiments demonstrate that our extended GAN setup yields more realistic data, accurately captures class imbalanced combinations found in clinical variables, maintains robust security, and ensures high utility. In essence, our method presents a solution to the restricted access to healthcare data, underscoring the potential of synthetic datasets for benchmarking machine learning models and healthcare education.

2. Background

The Difficulties of Training GANs

GANs, unlike most generative models [17–19], do not directly compute the data likelihood. Instead, they model complex probability distributions through a pair of sub-networks: a generator and a discriminator. By synthesising data from a random latent prior, the generator aims to generate data that can fool the discriminator into mistaking it for real data. Meanwhile, the discriminator aims to distinguish real data from synthetic data by maximising the difference between their distributions. GANs have achieved remarkable success in image generation [20,21] and have found applications in natural language processing [22,23]. However, their application in the medical domain remains under-explored; and MedGAN [24,25], a promising approach for generating synthetic medical data, has been found to lack diversity when representing multivariate categorical cancer data [26].[1]

Despite their potential, GANs suffer from practical challenges during training. In particular, mode collapse [15] causes the generator to output the same family of data hence reducing diversity. GANs are also notoriously difficult to train, and fine-tuning the sub-networks using gradient descent techniques (*e.g.,* SGD [27] and Adam [28]) may not result in convergence. Numerous improvements have been proposed, including careful selection of network modules [29], changing learning objectives [12,13], and auxiliary experimental setups [30,31], but these methods do not necessarily enhance the quality of the generated data.

Prior work have shown that enforcing the Lipschitz constraint on the discriminator/critic network can improve the quality of synthetic images [13,32]. However, allowing mode collapse can allocate more expressive power to fine-tune the few identified modes [33]. To mitigate mode collapse, previous studies have leveraged the learned features within the GAN sub-networks to quantify diversity in the generated data. One approach is to use *minibatch discrimination* [30], while another involves adopting an encoder–decoder framework for the discriminator [34]. Inspired by this line of research, our novel extended GAN setup (see Fig. 1) stores latent features of real data in an external buffer and replays them to the generator as a form of non-randomised prior at test time.

Another line of study also showed that the generator outputs can benefit from employing discriminators with better designs or by having multiple discriminators [35,36]. Thanh-Tung and Tran [37] found that mode collapse could be related to *catastrophic forgetting* [38,39] in the discriminator when the discriminator parameters escaped their previous local minimum. Mangalam and Garg [40] found that this could be mitigated by sequentially introducing more discriminators.

On Concurrently Modelling Mixed-Type Variables

While Goncalves et al. [26] showed that the traditional GAN approach encountered difficulties in generating multivariate categorical data, real life clinical data are even more complicated and usually consist of numeric, binary, and categorical variables. Two recent studies, Li

[1] Refer to Table 7 and Fig. 7 on page 16 of Goncalves et al. [26].

Table 1
The variables of the ART in HIV dataset.

Variable name	Data type	Unit	Valid categorical options
Viral Load (VL)	numeric	copies/mL	- -
Absolute Count for CD4 (CD4)	numeric	cells/μL	- -
Relative Count for CD4 (Rel CD4)	numeric	cells/μL	- -
Gender	binary	- -	Male; Female
Ethnicity (Ethnic)	categorical	- -	Asian; African; Caucasian; Other
Base Drug Combination (Base Drug Combo)	categorical	- -	FTC + TDF; 3TC + ABC; FTC + TAF; DRV + FTC + TDF; FTC + RTVB + TDF; Other
Complementary INI (Comp. INI)	categorical	- -	DTG; RAL EVG; Not Applied
Complementary NNRTI (Comp. NNRTI)	categorical	- -	NVP; EFV RPV; Not Applied
Extra PI	categorical	- -	DRV; RTVB; LPV; RTV; ATV; Not Applied
Extra pk Enhancer (Extra pk-En)	binary	- -	False; True
VL Measured (VL (M))	binary	- -	False; True
CD4 (M)	binary	- -	False; True
Drug Recorded (Drug (M))	binary	- -	False; True

et al. [41] and Kuo et al. [14], both reported the successful generation of mixed-type datasets using GANs.

Li et al. proposed a generator with a pair of VAEs. The VAEs map clinical variables of different types to a common feature representation space — one VAE encoded the numeric variables, and the other encoded the non-numeric variables. Li et al. further included a matching loss to minimise the distance in the representation pairs. Kuo et al.'s Health Gym GANs included a soft-embedding [42] algorithm to create a small size lookup table for each binary and categorical variable. This enables features of all types to be concatenated and simultaneously processed by the network.

3. Materials and methods

This section discusses the ground truth ART for HIV dataset, provides more details on Kuo et al.'s [14] Health Gym GANs, and introduces our extended GAN setup that mitigates mode collapse.

3.1. The ground truth ART for HIV dataset

We selected a cohort of 8916 people (with 332,800 records) from the EuResist database [43] using published inclusion/exclusion criteria [14]. There are 3 numeric, 5 binary, and 5 categorical variables, as listed in Table 1. The numeric variables – VL, CD4, and Rel CD4 – are indicative of the patient's health status. The HIV treatment regimens [44] are deconstructed as Base Drug Combo, Complimentary (Comp.) INI, Comp. NNRTI, Extra PI, and Extra pk-En. The medication classes are: nucleoside reverse transcriptase inhibitors (NRTIs), nucleotide reverse transcriptase inhibitors (NtRTIs), non-nucleotide reverse transcriptase inhibitors (NNRTIs), integrase inhibitor (INI), and protease inhibitors (PIs). The base drug combo mainly comprises NRTIs + NtRTIs; and there are 50 medication combinations spanning 21 medications.

The original EuResist database also contains a considerable proportion of missing data. Missingness in clinical data is often highly informative [45], indicating *e.g.,* the need for specific laboratory tests. We include the variables with suffix (M) to indicate if measurements are taken. Measurements are taken (*i.e.,* the variable is set to True) in 24.27% of timepoints for VL (M), 22.21% for CD4 (M), and 85.13% for Drug (M) in the real dataset.

Data in EuResist are collected irregularly hence we summarised it across calendar months (taking the last reported value for each variable of that month). There are often long gaps (over 6 months) in the original records, hence we split such records into multiple shorter sub-records. We truncate the sub-records' lengths to the closest multiple of ten; as a result, the shortest record has 10 months of data and the longest has 100 months. Time-series data including medication usage are valuable for developing algorithms (*e.g.,* RL) that optimise and adapt treatment type and dosage over time.

3.2. The highly sparse and strongly correlated nature of clinical datasets

In Section 2, we discussed the issue of GAN-based models being unable to accurately represent multivariate medical data, as highlighted in Goncalves et al. [26]. To better understand this issue, we examine the context of ART for HIV treatment, which is characterised by frequent changes in medication regimes to prevent the development of drug-resistant viral strains, as reported in Bennett et al. [46] and Chapter 4 of World Health Organisation [47].

Consider the following 3 common medication combinations[2] for adolescents:

- `Option A`: TDF + FTC (Backbone) + `EFV` (Comp. NNRTI) + `N/A` (Comp. INI)
- `Option B`: TDF + FTC (Backbone) + `NVP` (Comp. NNRTI) + `N/A` (Comp. INI)
- `Option C`: TDF + FTC (Backbone) + `N/A` (Comp. NNRTI) + `DTG` (Comp. INI)

These examples include tenofovir (TDF), emtricitabine (FTC), efavirenz (EFV), nevirapine (NVP), and dolutegravir (DTG). When we change the medications `OPTION A`→`OPTION B`, the comp. NNRTI medication of `EFV` is replaced with the comp. NNRTI medication of `NVP`. When alternatively we change `OPTION A`→`OPTION C`, `EFV` is replaced with the comp. INI medication of `DTG`. There are sparsity, denoted as `N/A`, when an NNRTI medication is used instead of an INI medication and vice versa.

[2] See Table 4.3 on page 106 in World Health Organisation [47].

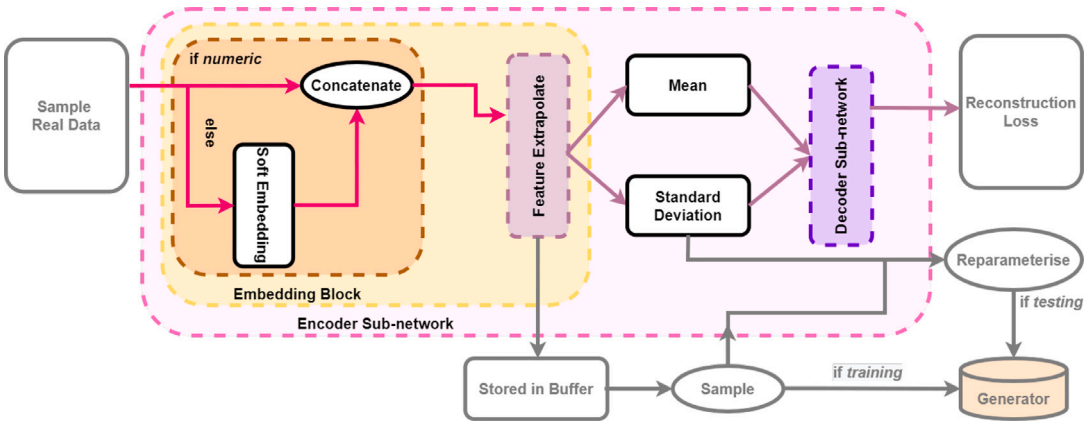


Fig. 2. The VAE is built into the critic to collect and extract real data features.

The sparsity results in the emergence of strong negative correlations among different variables due to the mutual exclusivity of these medications. We hypothesise that the highly sparse and strongly correlated nature of multivariate categorical clinical datasets can exacerbated mode collapse. Since `Options A`, `B`, and `C` are all realistic, GAN can potentially learn to only output EFV whenever the comp. INI medication class is `N/A`. This is further complicated when the variable distributions are highly skewed (*e.g.,* when EFV is prescribed more often than NVP), or when the distributions among different variables are imbalanced (*e.g.,* when NNRTI medication is prescribed more often than INI medication).

3.3. The Health Gym GAN

We based our work on Kuo et al.'s [14] Health Gym GAN. Their *generator* G and *critic* C were trained to optimise the losses,

$$L_C = \mathbb{E}\left[C\left(G(z)\right)\right] - \mathbb{E}\left[C(x_{\text{real}})\right] + \lambda_{\text{GP}}\mathbb{E}\left[\left(\left(\|\nabla_{\tilde{x}_{\text{syn}}} C(\tilde{x}_{\text{syn}})\|_2 - 1\right)^2\right]\right] \quad \text{and} \tag{1}$$

$$L_G = -\mathbb{E}\left[C(G(z))\right] + \lambda_{\text{corr}} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \text{abs}\left(r_{\text{syn}}^{(i,j)} - r_{\text{real}}^{(i,j)}\right), \tag{2}$$

following the *Wasserstein GAN with gradient penalty* (WGAN-GP) setup[3] [13]. $G(z) = x_{\text{syn}}$ denotes the synthetic data generated from the generator with random input z; x_{real} is the real data sampled from the database; $\tilde{x}_{\text{syn}}$ is interpolated between $G(z)$ and x_{real}; and λ_{GP} is a constant that manages the strength of the gradient penalty loss. Furthermore, Kuo et al. introduced an *alignment loss* (*i.e.,* the second term) in Eq. (2), computed as the sum of the absolute differences in the *Pearson's r correlation* [48] for all pairs of distinct variables $i \neq j$ between the synthetic data $r_{\text{syn}}^{(i,j)}$ and the real data $r_{\text{real}}^{(i,j)}$, with a constant λ_{corr} managing the strength.

3.4. Extension: Preconditioning the generator input using feature replay

We expand on Kuo et al.'s [14] WGAN-GP by incorporating a VAE to generate diverse synthetic patient records. As depicted in Fig. 1, we encode features from the data using the VAE and replay the features to the generator as a form of non-random inputs. In Fig. 2, we show that the replay mechanism is built in the critic by incorporating the VAE. When real data is passed through the critic, the VAE encodes the data as features γ_{real} with learned standard deviations σ_V.

[3] The critic in the WGAN setting [12] is equivalent to the discriminator of the traditional GAN setup. It is called a critic since it is not trained to classify outputs but to score their realisticness instead.

During training, we employ an external buffer $\mathscr{B}$ to collect γ_{real}. To ensure that our algorithm does not incur a high memory cost, our external buffer is set to a fixed size. When no vacancy is left in $\mathscr{B}$, we randomly release space in $\mathscr{B}$ to append the new encoded features γ_{real}. There are alternative ways to update the buffer, such as *herding* [49] for constructing an exemplar set [50]; but the search for an optimal buffer update mechanism is out of the scope of this paper.

At test time, we discard the VAE and sample stored features (with replacement) from the buffer for the GAN generator $\overline{\gamma_{\text{real}}} \sim \mathscr{B}$. The goal of the feature replay mechanic is to avoid mode collapse by establishing a dependency between the generator output and the highly diversified ground truth latent features of the real data γ_{real}. Novel synthetic patient records are created by reparameterising the generator input as $z \leftarrow \overline{\gamma_{\text{real}}} + \rho^*$ where $\rho^* \sim N(0, \sigma_V)$.

We discuss further implementation details of Sections 3.3 and 3.4 in § A of the Supplementary Materials. The supplement includes pseudocode outlining the training process of our augmented framework and how we collect (and release) the encoded data features.

4. Experimental setups

4.1. Hyper-parameters

We extend the architecture of our model based on Kuo et al.'s [14] WGAN-GP and inheriting most of their setups. Both the generator G and critic C utilise bi-directional *long short-term memory* (bi-LSTM) [51,52], a recurrent neural network (RNN) adept at handling sequential data, a key element in generating the ART for HIV dataset.

We employ soft-embedding (see Section 2) to transform binary and categorical variables into continuous vectors for feature learning. Furthermore, a VAE V forms an integral part of the critic C in our extended architecture. The encoder of V consists of a series of linear transformations and leverages *residual connections* [53] to enhance feature extrapolation. The decoder of V is handled by a single linear layer.

The Adam optimiser [28] is employed in training the model components of G, C, and V. To adeptly handle the inherent variability in length of medication records present in the ART for HIV dataset, we adopt a strategy of *curriculum learning* [54]. The model is initially trained on simpler tasks (*i.e.,* the synthesis of shorter records), before it is gradually exposed to increasingly complex tasks (*i.e.,* the synthesis of longer records). This phased progression is crucial in optimising model performance.

For a more exhaustive exposition on the module and optimisation configurations, we direct the readers to Section § B.1 in the Supplementary Materials.

4.2. Baseline models

Our proposed model is referenced as WGAN-GP+VAE+Buffer, while the initial setup proposed by Kuo et al. [14] is denoted as WGAN-GP.

Alternative Architectural Designs

The performance of our models can be significantly influenced by architectural choices. For instance, we found that replacing LSTM units with Transformers [55] in the WGAN-GP model led to improved results. Among the Transformer variants we experimented with, including vanilla Transformers, BERT-like encoder-only Transformers [56], and GPT-like decoder-only Transformers [57], we found the encoder-only Transformers to be the most beneficial. We denote this new setup as WGAN-GP+G_EOT+VAE+Buffer. For comprehensive specifications of the Transformer setup, please refer to Section § B.2.1 in the Supplementary Materials.

Previous Methods to Mitigate Mode Collapse

The WGAN-GP model was further extended using three techniques aimed at reducing mode collapse (see Section 2). These techniques include minibatch discrimination (MBD) [30], moment matching (MM) [34], and multiple critics (MC) [40], resulting in the WGAN-GP+MBD, WGAN-GP+MM, and WGAN-GP+MC models, respectively. Detailed implementation of these extensions is provided in Section § B.2.2 in the Supplementary Materials.

Additionally, we experimented with a design inspired by the VAE-GAN model [58], referred to as VAE-WGAN-GP. This model utilises a VAE encoder in addition to the original Kuo et al. setup. A detailed comparison of our approach with Larsen et al.'s model can be found in Section § B.3 of the Supplementary Materials.

4.3. Evaluation metrics

In assessing our synthetic dataset, we uphold five essential criteria: (a) mitigation of mode collapse in GAN generator; (b) generation of individually realistic variables; (c) the collective realism of variables over time; (d) assurance of privacy; and (e) utility for developing downstream machine learning algorithms. For a thorough comprehension of these criteria, we propose an array of relevant metrics.

4.3.1. Mitigating mode collapse

To evaluate the presence of mode collapse, we leverage two metrics — the *log-cluster* metric U [59] and the *category coverage* (CAT) metric [26]. These metrics respectively quantify the differences in latent structures between the real and synthetic datasets, and the completeness of non-numeric classes in the synthetic datasets. Computational details on these metrics can be found in § B.4.1 of the Supplementary Materials.

4.3.2. Realism of individual variables

We visually assess the realism of individual variables using *kernel density estimations* (KDEs) [60] for numeric variables and side-by-side bar plots for categorical variables. Additionally, we employ the two-sample Kolmogorov–Smirnov test [61], the two independent sample Student's t-test [62], the F-test [63], and the three sigma rule test [64]. These statistical tests ensure that our synthetic variables closely resemble their real counterparts. More information on these tests can be found in § B.4.2 of the Supplementary Materials.

4.3.3. Preserving variable correlations

To evaluate the fidelity of variable correlations, we calculate Kendall's τ rank correlation [65] for the mixed-type datasets. We consider both *static* correlations among all data points and average *dynamic* correlations, following the method in Kuo et al. [14]. Refer to more descriptions in § B.4.3 of the Supplementary Materials.

4.3.4. Privacy assurance

To ensure privacy, we conduct two tests to prevent the disclosure of private patient information. We ascertain that no real record is leaked into the synthetic dataset, and we apply El Emam et al. [1]'s *sample-to-population attack* to examine whether new information can be learned by matching an individual from the synthetic dataset to the real dataset. More information on the procedures can be found in § B.4.4 of the Supplementary Materials.

4.3.5. Utility verification

The utility of our synthetic dataset is evaluated by training RL agents on both real and synthetic datasets [66,67]. High utility is achieved when these agents suggest similar actions for patient treatment, demonstrating that the synthetic dataset can be an effective substitute for the real dataset in downstream machine learning applications.

For a complete description of our evaluation metrics, including the corresponding mathematical formulae, please refer to § B.4.5 in the Supplementary Materials. By using these metrics, we believe that our synthetic dataset not only emulates the real dataset closely but also meets the requirements of privacy preservation and practical utility [68].

5. Results

This section evaluates the five desiderata outlined in Section 4.3.

5.1. Mitigating mode collapse

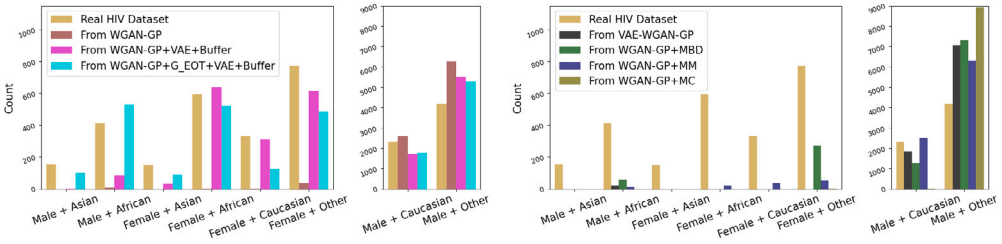
We trained seven variants of GANs and compared the generated synthetic datasets using the metrics discussed in Section 4.3.1. We found that our proposed approach including an extra VAE with external buffer was better at mitigating mode collapse. They score −3.00 in the log-cluster scores, lower than the −2.11 by Kuo et al.'s [14] baseline model; indicating that our synthetic datasets $\mathfrak{D}_{\text{alt}}$ are better at mirroring the latent structure in the ground truth $\mathfrak{D}_{\text{real}}$. In addition, both of our $\mathfrak{D}_{\text{alt}}$ score 1.00 in category coverage, thus they cover all categories that can be found in $\mathfrak{D}_{\text{real}}$. Refer to § C.2 in the Supplementary Materials for all quantitative scores.

The quantitative results can be contextualised with the qualitative investigation shown in Fig. 3. We plotted the frequency of all *gender-ethnicity pairs*; subplot 3(a) shows that Kuo et al. [14]'s generated data $\mathfrak{D}_{\text{null}}$ suffers from mode collapse and includes mostly *Male+Caucasian* and *Male+Other* patients, and omitted patients of minority background (*e.g.,* Asian males and African females). In contrast, our $\mathfrak{D}_{\text{alt}}$ captured both genders for all ethnicities. We also found that previous methods that were proposed to mitigate mode collapse in computer vision were not very effective for clinical time-series. In subplot 3(b), the synthetic datasets generated using MBD [30], MM [34], and MC [40] all behaved similarly to (or worse than) Kuo et al.'s baseline.

5.2. Realisticness of the individual variables

Fig. 4 presents the KDE plots and side-by-side barplots for the individual variable comparisons. The real variables from $\mathfrak{D}_{\text{real}}$ are in gold; subplot 4(a) illustrates the synthetic variables in $\mathfrak{D}_{\text{null}}$ generated using WGAN-GP [14] in brown, and subplot 4(b) depicts those in $\mathfrak{D}_{\text{alt}}$ simulated with our WGAN-GP+G_EOT+VAE+Buffer in cyan.

Overall, the distributions in subplot 4(b) are more similar than those in subplot 4(a). Thus our dataset $\mathfrak{D}_{\text{alt}}$ captures more details in the ground truth $\mathfrak{D}_{\text{real}}$. Specifically, our $\mathfrak{D}_{\text{alt}}$ is more capable of representing females in gender, Asians in ethnicity, and the prescription of less common medications (*e.g.,* NVP in the NNRTI medication class and DRV in the PI medication class). In contrast, $\mathfrak{D}_{\text{null}}$ exhibits a bias towards the dominant class in the binary and categorical variables (*e.g.,* making the medication in $\mathfrak{D}_{\text{null}}$ even more unlikely to include Extra pk-En than



(a) Real dataset vs our synthetic dataset candidates. (b) Other synthetic dataset candidates.

Fig. 3. Inspecting mode collapse within the gender-ethnicity pair.

that in $\mathfrak{D}_{\text{real}}$). This hence shows that the additional VAE and extra buffer in our setup was effective in capturing extreme class imbalanced distributions in real world clinical data.

We then examined the synthetic variables using a series of hierarchically structured statistical tests outlined in Section 4.3.2. All variables of the WGAN-GP+G_EOT+VAE+Buffer synthetic dataset passed the KS test revealing that all variables of our synthetic dataset are realistic and capture both the mean and variance of their real counterparts. In contrast, the synthetic VL distribution in Kuo et al.'s WGAN-GP generated dataset failed the KS test because it was not able to mirror the spread of the real VL distribution.[4] Refer to § C.5 in the Supplementary Materials for the complete statistical outcomes and extra results on our WGAN-GP+VAE+Buffer setup.

5.3. Correlations among variables

Following Section 4.3.3, we checked the fidelity among variable pairs and presented the correlations among the variables in Fig. 5. Overall, Kuo et al.'s [14] WGAN-GP and our two extended setups were all able to generate synthetic datasets with realistic correlations. This applied to both the static correlation in subplot 5(a) and the dynamic correlations in subplots 5(b) and (c).

We noticed that the variables in $\mathfrak{D}_{\text{null}}$ generated using the baseline WGAN-GP have the tendency to increase the magnitudes of correlations. Some examples of this behaviour can be found in the pairs of (Drug (M), CD4) and (Extra PI, Base Drug Combo) in the dynamic correlation in trends; and likewise for (CD4 (M), VL) in the dynamic correlations in cycles.

In contrast, it could be argued that the correlations between variables in $\mathfrak{D}_{\text{alt}}$ generated using our WGAN-GP+G_EOT+VAE+Buffer tend to be weaker than those in the real dataset. This is observed in (CD4, VL) in the dynamic correlation in trends and similarly in (CD4 (M), VL (M)) in the dynamic correlation in cycles.

5.4. The patient disclosure risk

Since our aim is to create realistic synthetic data available for public access, we evaluated the risk of patient re-identification as discussed in Section 4.3.4. The minimum Euclidean distance between the real dataset and the synthetic dataset generated using WGAN-GP+VAE+Buffer is 0.1029. It is 0.1229 with the synthetic dataset generated via WGAN-GP+G_EOT+VAE+Buffer. Hence no real record is leaked into the synthetic dataset using either of our setups. Note, it is not meaningful to compare the magnitudes of the Euclidean distances and what we desire is that they are > 0.

Using El Emam et al.'s [1] metric, the disclosure risk of WGAN-GP+VAE+Buffer's synthetic dataset is 0.044% while it is 0.095% for WGAN-GP+G_EOT+VAE+Buffer's synthetic dataset. All of the results are much lower than the 9% threshold suggested by Health Canada [69] and the European Medicines Agency [70]. Thus, combined with our prior results in this section, we conclude that our synthetic ART for HIV datasets generated using our extended WGAN-GP setup are both realistic and secure.

5.5. Data utility

Following Section 4.3.5, we tested the utility of the synthetic datasets by training RL agents to optimise ART medication combinations. We visualised the relative frequencies of the actions taken by the trained RL agents using heatmaps. Each tile represents a unique action, and the number on the tile represents the frequency of that action, as a proportion of all actions. This section primarily focuses on the synthetic dataset generated using WGAN-GP+G_EOT+VAE+Buffer.

5.5.1. General utility

Fig. 6 illustrates the actions taken by the RL agents when the action space was spanned by Comp. NNRTI and Base Drug Combo. Subplot (a) presents the actions taken by an RL agent trained on the real dataset $\mathfrak{D}_{\text{real}}$; subplot (b) shows the RL agent trained on the synthetic dataset $\mathfrak{D}_{\text{null}}$ generated by WGAN-GP [14]; and likewise subplot (c) shows an RL agent trained on the synthetic dataset $\mathfrak{D}_{\text{alt}}$ generated using our WGAN-GP+G_EOT+VAE+Buffer.

The heatmap in subplot (b) does not match the one in subplot (a), indicating that the RL agent trained on $\mathfrak{D}_{\text{null}}$ was incapable of suggesting similar actions to the RL agent trained on $\mathfrak{D}_{\text{real}}$. The RL agent trained on $\mathfrak{D}_{\text{null}}$ suggested NVP for Comp. NRTI with DRV + FTC + TDF for Base Drug Combo for 48.97% of all its actions. The undiversified policy was likely induced by mode collapse in WGAN-GP – causing $\mathfrak{D}_{\text{null}}$ to capture only a fraction of all prescribed treatments.

In contrast, subplot (c) shows that the RL agent trained on our $\mathfrak{D}_{\text{alt}}$ exhibited a higher diversification in its treatment strategy. The heat map in subplot (c) mirrored the one in subplot (a) better, showing that $\mathfrak{D}_{\text{alt}}$ possesses a higher utility than the baseline $\mathfrak{D}_{\text{null}}$. In § C.8 of the Supplementary Materials, we test a similar scenario in which the action space is spanned by Comp. INI and Base Drug Combo.

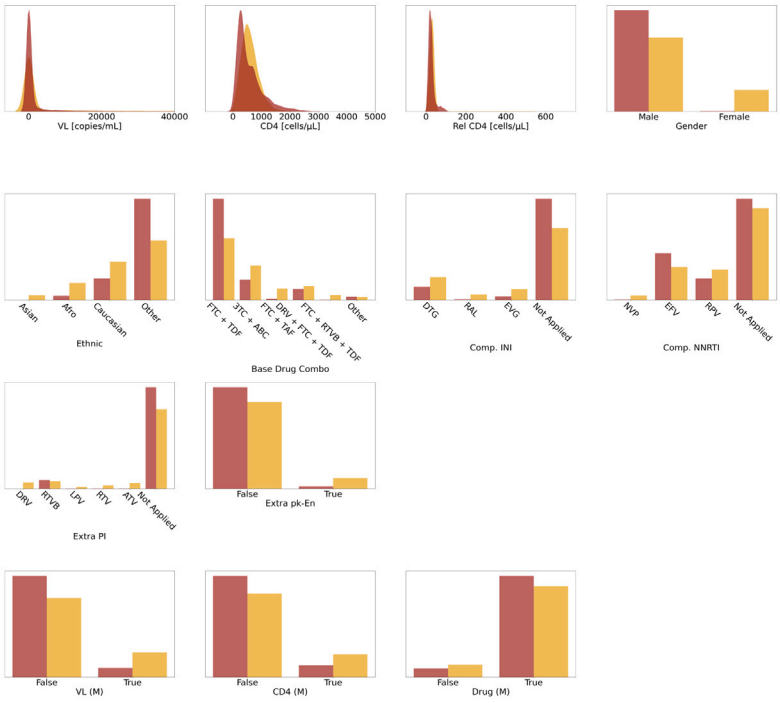
5.5.2. Utility in minority groups

Mode collapse in GANs has a particularly negative impact on downstream model utility for minority groups. To demonstrate the severity of this problem, we repeated the experiments in Section 5.5.1 but included only patients of African ethnicity. The results are shown in Fig. 7.

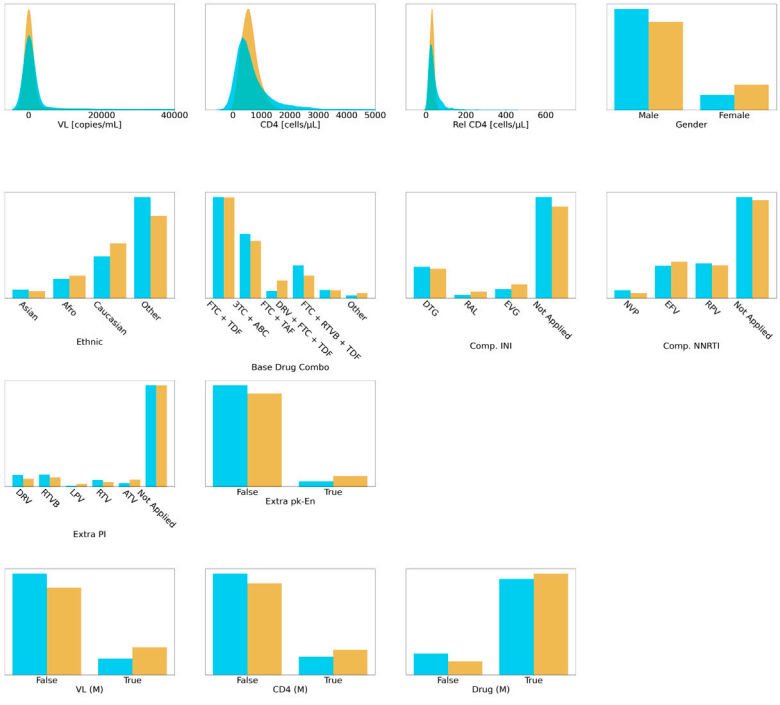
As previously demonstrated in Fig. 3, Kuo et al.'s [14] WGAN-GP experienced mode collapse and had difficulties in generating synthetic patients of Asian and African ethnicity. This meant that the number of data points in $\mathfrak{D}_{\text{null}}$ were insufficient to cover the diversity in ART for HIV regimens for this patient population. As a result, subplot 7(b) differs greatly from subplot 7(a) and the RL agent trained on $\mathfrak{D}_{\text{null}}$ was incapable of suggesting similar actions to the RL agent trained on $\mathfrak{D}_{\text{real}}$.

In stark contrast, subplot 7(c) captures most features that can be found in subplot 7(a). This shows that our synthetic dataset $\mathfrak{D}_{\text{alt}}$ has high utility and is more suitable to replace the baseline $\mathfrak{D}_{\text{null}}$ for

[4] See Table 7 on page 26 of Kuo et al. [14].



(a) Synthetic dataset $\mathfrak{D}_{\text{null}}$ from WGAN-GP (Kuo et al., 2022) in brown.



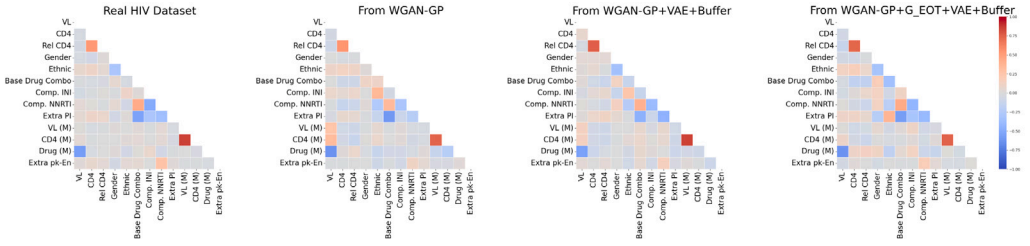
(b) Synthetic dataset $\mathfrak{D}_{\text{alt}}$ from WGAN-GP+G_EOT+VAE+Buffer (ours) in cyan.

Fig. 4. Comparing the individual variable distributions, with the real dataset in colour gold.

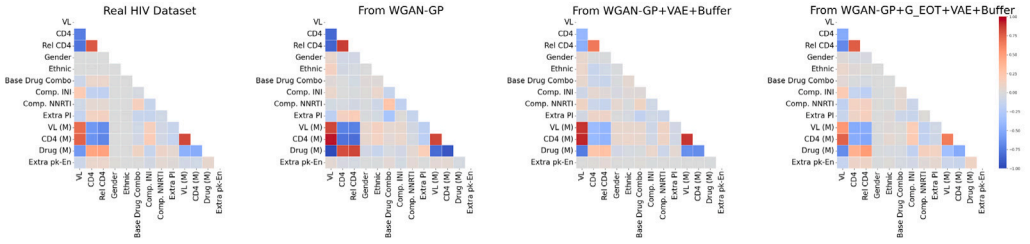
supporting the development of downstream machine learning algorithms. As discussed in Section 5.1, this was due to the effectiveness of WGAN-GP+G_EOT+VAE+Buffer in mitigating mode collapse when generating real world clinical data.

6. Discussion

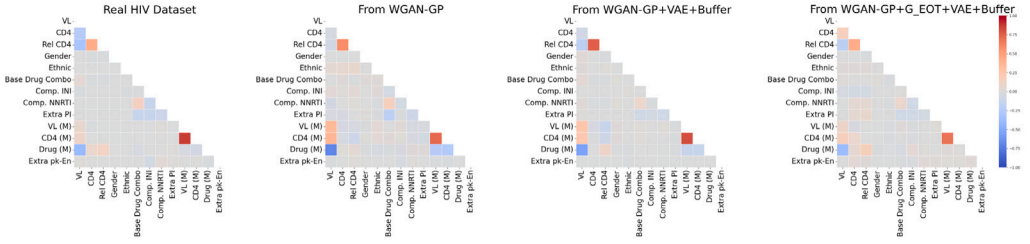
The creation of open, privacy-preserving datasets that adequately represent priority populations has been proposed as a way to reduce known disparities in health care [71]. GANs have previously been demonstrated to be able to generate highly realistic synthetic longitudinal patient records [14] . However, use of these datasets for developing artificial intelligence tools for healthcare might perpetuate health disparities if minority patient groups are poorly represented. We sought to tackle the problem of mode collapse in GANs through employing a VAE and buffer within a GAN structure.



(a) The classic static correlation.



(b) The dynamic correlation in trends.



(c) The dynamic correlation in cycles.

Fig. 5. Comparing different types of correlations in the real and synthetic datasets.

Our results demonstrated that the augmented approach, WGAN-GP+G_EOT+VAE+Buffer, is proficient at capturing the distinct imbalance in class distributions. Experimental comparisons between our model and the WGAN-GP method of Kuo et al. [14] showed superior ability in mimicking the original data distributions, particularly in representing both majority (*e.g.,* `Male Caucasian` and `Male Other`) and minority demographics (*e.g.,* `Male Asian` and `Male African`; see Section 5.1). Notably, our augmented structure retains the merits of Kuo et al.'s initial model, generating synthetic datasets with realistic correlations (see Section 5.3) and robust security metrics (see Section 5.4).

Our improved GAN-based method enhances both the *resemblance* and *representation* of synthetic data in relation to the real counterparts [72]. The former refers to the extent to which synthetic data match the real data; and the latter considers the adequacy of coverage of priority populations within the generated data. These concepts align with the ethical guidelines in artificial intelligence stated in the Australian Government Department of Industry, Science, and Resources [73] and the United States of America Government Food and Drug Administration [74]. Thus, our proposed model represents a promising resource for generating highly secure, realistic synthetic data for public use. Synthetic data can be used for educational purposes, to accelerate the development of machine learning models in health care, and for the testing and development of clinical software systems.

Our primary contribution lies in demonstrating the heightened utility of our synthetic data, simulated using WGAN-GP+G_EOT+VAE+Buffer, particularly for patients of `African` ethnicity, when compared to the synthetic data produced by Kuo et al.'s WGAN-GP. Here, *utility* refers to the ability of a synthetic dataset to effectively replace the real dataset for developing downstream machine learning algorithms.

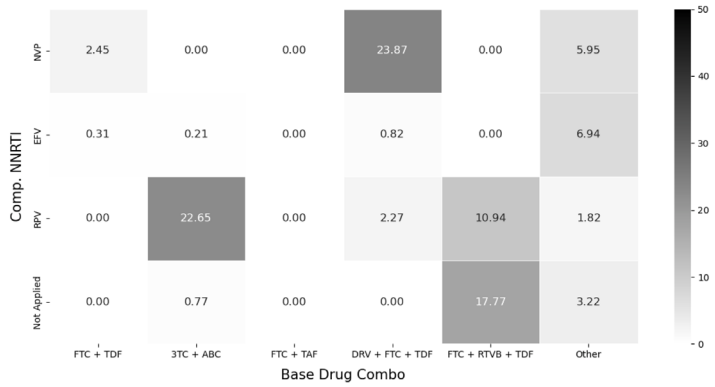
In Section 5.5, we showed that RL agents trained on our synthetic dataset $\mathfrak{D}_{\text{alt}}$ yielded a policy that more closely parallelled the policy learned from the real dataset $\mathfrak{D}_{\text{real}}$, in comparison to the policy derived from Kuo et al.'s synthetic dataset $\mathfrak{D}_{\text{null}}$. This highlights the limitations of the conventional GAN model used by Kuo et al. in accurately capturing the nuances of minority patient data.

Although these findings are encouraging, we remark that we have utilised a single type of RL algorithm for our utility testing (see Section § B.4.5 in the Supplementary Materials). Future work will extend our evaluation to include a wider range of offline RL techniques [75] and traditional statistical methods, such as decision trees [76]. Further research could also compare our work with other generative models, such as diffusion models [77,78]. We are confident that continuous exploration of this kind will augment synthetic data generation techniques and contribute to the progress of robust, equitable open data for artificial intelligence and machine learning in healthcare.

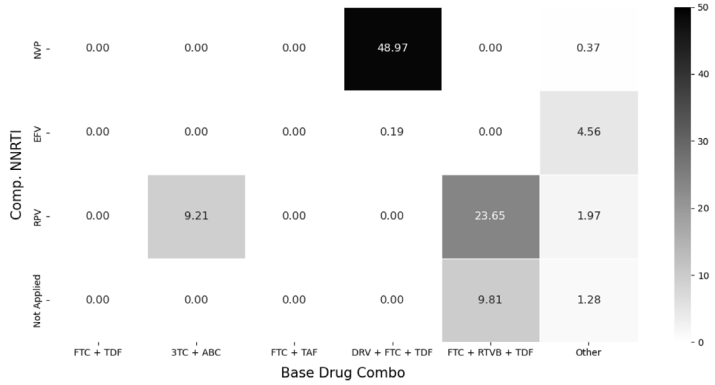
7. Conclusion

We devised a novel solution to tackle mode collapse, a significant challenge in synthesising clinical time-series of mixed-typed data using GANs. Our methodology employs a VAE and buffer within a GAN structure, effectively serving as a memory store to preserve key features from real patient data during training. These retained details inform the GAN in enhancing synthetic data with higher diversity and mitigating mode collapse.

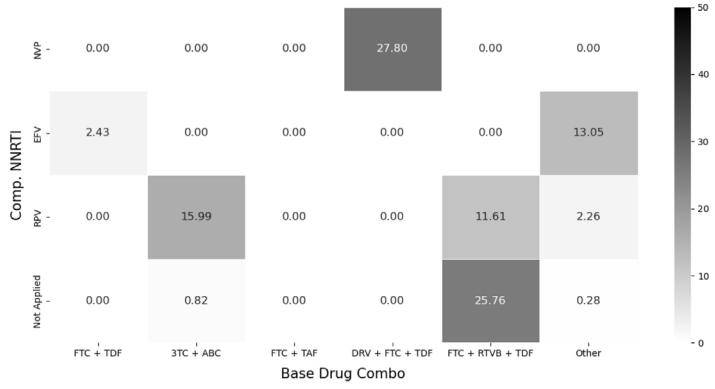
Our method demonstrated promising results on the ART for HIV dataset from the EuResist database, outperforming existing methods in generating a diverse patient cohort, with a particular emphasis on accurately representing minority groups. This attribute is crucial in biomedical informatics, where diverse and representative data are required for fair and effective decision-making algorithms.



(a) Trained using the real dataset $\mathfrak{D}_{\text{real}}$.



(b) Trained using the synthetic dataset $\mathfrak{D}_{\text{null}}$ generated by WGAN-GP (Kuo et al., 2022).



(c) Trained using the synthetic dataset $\mathfrak{D}_{\text{alt}}$ generated by our WGAN-GP+G_EOT+VAE+Buffer.

Fig. 6. The suggestions made by RL agents trained on different ART for HIV datasets with Comp. NNRTI and Base Drug Combo spanning the action space.

Data access

Our synthetic dataset, generated using WGAN-GP+G_EOT+VAE+Buffer, can be found on our website https://healthgym.ai/.[5] Spanning 60 months and encompassing 8916 patients, the dataset totals 534,960 records (= 8916 × 60). Each record features 15 columns, inclusive of the 13 ART for HIV variables as outlined in Table 1, a patient identifier, and a time point specification.

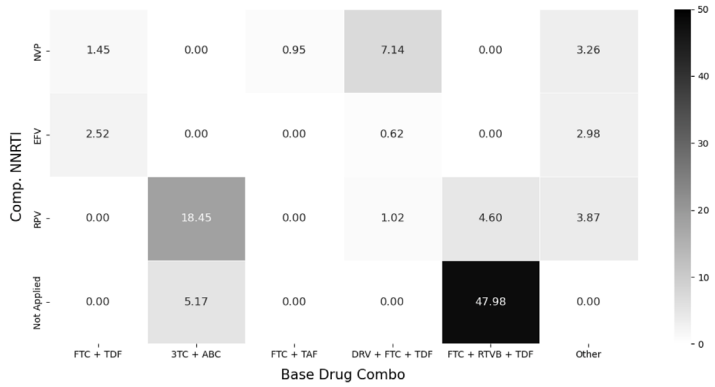
While our method was applied to the ART for HIV dataset, its potential extends to other clinical time-series datasets, especially those with imbalanced categorical variables. The synthetic datasets we create, although not replacements for real data, offer valuable resources for machine learning development, especially where privacy is paramount. Our synthetic dataset is publicly available for researchers and practitioners in biomedical informatics.[6]

Ethics statement & reproducibility

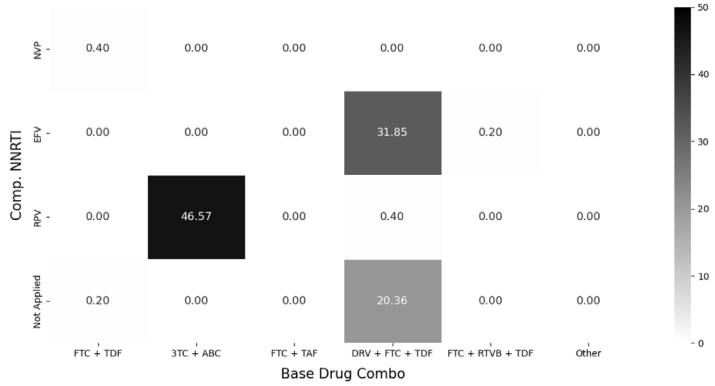
This study was approved by the University of New South Wales' human research ethics committee (application HC210661). We based our synthetic HIV dataset on EuResist [43]. For people in the EuResist integrated database, all data providers obtained informed consent for the execution of retrospective studies and inclusion in merged cohorts [80].

[5] See: Kuo [79] on Figshare, the research-oriented open access repository.

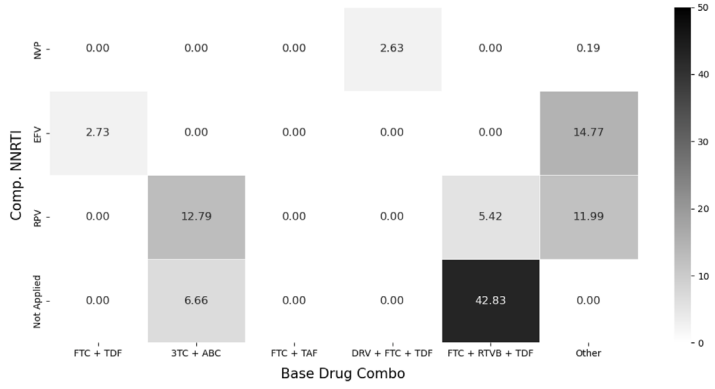
[6] Our dataset was also used in an event to foster collaboration between academics, clinicians, and health facilities in New South Wales, Australia. See: https://cbdrh-hds-datathon-2023.github.io/.



(a) Trained using the real dataset $\mathfrak{D}_{\text{real}}$.



(b) Trained using the synthetic dataset $\mathfrak{D}_{\text{null}}$ generated by `WGAN-GP` (Kuo et al., 2022).



(c) Trained using the synthetic dataset $\mathfrak{D}_{\text{alt}}$ generated by our `WGAN-GP+G_EOT+VAE+Buffer`.

Fig. 7. The suggestions made by RL agents trained on African patients in different ART for HIV datasets with Comp. NNRTI and Base Drug Combo spanning the action space.

The EuResist Integrated DataBase (EIDB) can be accessed for scientific studies once a proposal for analysis has been approved by EuResist's Scientific Board (see: http://engine.euresist.org/database/). To facilitate future research, our code will be made available after the paper is published. Our synthetic dataset is freely accessible through https://healthgym.ai/.

Problem:	Clinical data is highly confidential and cannot be freely distributed, which hampers the development of machine learning models in healthcare.
What is Already Known:	Generative adversarial networks (GANs) can generate realistic synthetic datasets, but they suffer from mode collapse which reduces cohort diversity and results in sub-optimal performance for minority populations.
What this Paper Adds:	We extended the classic GAN setup with an additional variational autoencoder and external memory, thereby overcoming mode collapse. This generates privacy-preserving synthetic datasets with accurate class distributions and high utility for prototyping and benchmarking machine learning models in healthcare.

CRediT authorship contribution statement

Nicholas I-Hsien Kuo: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Federico Garcia:** Resources. **Anders Sönnerborg:** Resources. **Michael Böhm:** Resources. **Rolf Kaiser:** Resources. **Maurizio Zazzi:** Resources. **EuResist Network study group:** Resources. **Mark Polizzotto:** Validation. **Louisa Jorm:** Conceptualization, Validation, Supervision, Project administration. **Sebastiano Barbieri:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nicholas I-Hsien Kuo reports financial support was provided by Wellcome Trust Open Research Fund. Sebastiano Barbieri reports financial support was provided by Wellcome Trust Open Research Fund.

Acknowledgements

This study benefited from data provided by EuResist Network EIDB; and this project has been funded by a Wellcome Trust Open Research Fund (reference number 219691/Z/19/Z).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104436.

References

[1] Khaled El Emam, Lucy Mosquera, Jason Bass, Evaluating identity disclosure risk in fully synthetic health data: Model development and validation, J. Med. Internet Res. 22 (2020) 23139.
[2] Rachel Nosowsky, Thomas J. Giordano, The health insurance portability and accountability act of 1996 (HIPAA) privacy rule: Implications for clinical research, Annu. Rev. Med. 57 (2006) 575–590.
[3] Christine M. O'Keefe, Chris J. Connolly, Privacy and the use of health data for research, Med. J. Aust. 193 (9) (2010) 537–541.
[4] Heidi Beate Bentzen, Rosa Castro, Robin Fears, George Griffin, Volker Ter Meulen, Giske Ursin, Remove obstacles to sharing health data with researchers outside of the European Union, Nat. Med. 27 (8) (2021) 1329–1333.
[5] Cheryl Jones, Brenda Gannon, Abel Wakai, Ronan O'Sullivan, A systematic review of the cost of data collection for performance monitoring in hospitals, Syst. Rev. 4 (2015) 1–10.
[6] Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F.K. Williamson, Faisal Mahmood, Synthetic data in machine learning for medicine and healthcare, Nat. Biomed. Eng. 5 (6) (2021) 493–497.
[7] Nan Sun, Jun Zhang, Paul Rimba, Shang Gao, Leo Yu Zhang, Yang Xiang, Data-driven cybersecurity incident prediction: A survey, IEEE Commun. Surv. Tutor. 21 (2) (2018) 1744–1772.
[8] Richard S. Sutton, Andrew G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.
[9] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, A. Aldo Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nature Med. 24 (2018) 1716–1720.
[10] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, Leo Anthony Celi, Guidelines for reinforcement learning in healthcare, Nature Med. 25 (2019) 16–18.
[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial nets, in: The Advances in Neural Information Processing Systems, 2014.
[12] Martin Arjovsky, Soumith Chintala, Léon Bottou, Wasserstein generative adversarial networks, in: The International Conference on Machine Learning, 2017, pp. 214–223.
[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron C. Courville, Improved training of Wasserstein GANs, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), The Advances in Neural Information Processing Systems, 2017.
[14] Nicholas I. Kuo, Mark N. Polizzotto, Simon Finfer, Federico Garcia, Anders Sönnerborg, Maurizio Zazzi, Michael Böhm, Rolf Kaiser, Louisa Jorm, Sebastiano Barbieri, et al., The health gym: Synthetic health-related datasets for the development of reinforcement learning algorithms, Sci. Data 9 (1) (2022) 1–24.
[15] Ian Goodfellow, NeurIPS 2016 tutorial: Generative adversarial networks, 2016, Preprint at https://arxiv.org/abs/1701.00160.
[16] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, Krasimira Tsaneva-Atanasova, Artificial intelligence, bias, and clinical safety, BMJ Qual. Saf. 28 (2019) 231–237.
[17] Diederik P. Kingma, Max Welling, Auto-encoding variational Bayes, in: The International Conference on Learning Representations, 2014.
[18] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, Surya Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: The International Conference on Machine Learning, 2015, pp. 2256–2265.
[19] Aaron Van Oord, Nal Kalchbrenner, Koray Kavukcuoglu, Pixel recurrent neural networks, in: The International Conference on Machine Learning, 2016, pp. 1747–1756.
[20] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang, Generative image inpainting with contextual attention, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5505–5514.
[21] Tero Karras, Samuli Laine, Timo Aila, A style-based generator architecture for generative adversarial networks, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.
[22] Jingjing Xu, Xuancheng Ren, Junyang Lin, Xu Sun, Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation, in: The Empirical Methods in Natural Language Processing, 2018, pp. 3940–3949.
[23] Santiago Pascual, Antonio Bonafonte, Joan Serrà, SEGAN: Speech enhancement generative adversarial network, Interspeech (2017) 3642–3646.
[24] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, Jimeng Sun, Generating multi-label discrete patient records using generative adversarial networks, in: The Machine Learning for Healthcare Conference, 2017, pp. 286–305.
[25] Ramiro Camino, Christian Hammerschmidt, Radu State, Generating multi-categorical samples with generative adversarial networks, in: The ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models, 2018.
[26] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, Ana Paula Sales, Generation and evaluation of synthetic patient data, BMC Med. Res. Methodol. 20 (2020) 1–40.
[27] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.
[28] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimisation, in: The International Conference on Learning Representations, 2015.
[29] Alec Radford, Luke Metz, Soumith Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, Preprint at https://arxiv.org/abs/1511.06434.
[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, Improved techniques for training GANs, in: The Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.
[31] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, Ferenc Huszár, Amortised map inference for image super-resolution, in: The International Conference on Learning Representations, 2016.
[32] Kanglin Liu, Wenming Tang, Fei Zhou, Guoping Qiu, Spectral regularisation for combating mode collapse in GANs, in: The IEEE International Conference on Computer Vision, 2019, pp. 6382–6390.
[33] Luke Metz, Ben Poole, David Pfau, Jascha Sohl-Dickstein, Unrolled generative adversarial networks, in: The International Conference on Learning Representations, 2016.
[34] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, Barnabás Póczos, MMD GAN: Towards deeper understanding of moment matching network, in: The Advances in Neural Information Processing Systems, 2017, pp. 2200–2210.
[35] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U. Gutmann, Charles Sutton, VEEGAN: Reducing mode collapse in GANs using implicit variational learning, in: The Advances in Neural Information Processing Systems, 2017, pp. 3310–3320.
[36] Gonçalo Mordido, Haojin Yang, Christoph Meinel, Microbatchgan: Stimulating diversity with multi-adversarial discrimination, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 3061–3070.
[37] Hoang Thanh-Tung, Truyen Tran, Catastrophic forgetting and mode collapse in GANs, in: The International Joint Conference on Neural Networks, 2020, pp. 1–10.
[38] Michael McCloskey, Neal J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of Learning and Motivation, vol. 24, 1989, pp. 109–165.
[39] Nicholas I Kuo, Mehrtash Harandi, Nicolas Fourrier, Christian Walder, Gabriela Ferraro, Hanna Suominen, Learning to continually learn rapidly from few and noisy data, in: The Meta-Learning and Co-Hosted Competition of the AAAI Conference on Artificial Intelligence, 2021, pp. 65–76.

[40] Karttikeya Mangalam, Rohin Garg, Overcoming mode collapse with adaptive multi adversarial training, 2021, Preprint at https://arxiv.org/abs/2112.14406.
[41] Jin Li, Benjamin J. Cairns, Jingsong Li, Tingting Zhu, Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications, 2021, Preprint at https://arxiv.org/abs/2112.12047.
[42] Alejandro Mottini, Alix Lheritier, Rodrigo Acuna-Agost, Airline passenger name record generation using generative adversarial networks, 2018, Preprint at https://arxiv.org/abs/1807.06657.
[43] Maurizio Zazzi, Francesca Incardona, Michal Rosen-Zvi, Mattia Prosperi, Thomas Lengauer, Andre Altmann, Anders Sonnerborg, Tamar Lavee, Eugen Schülter, Rolf Kaiser, Predicting response to antiretroviral treatment by machine learning: the EuResist project, Intervirology 55 (2) (2012) 123–127.
[44] Michele W. Tang, Tommy F. Liu, Robert W. Shafer, The HIVdb system for HIV-1 genotypic resistance interpretation, Intervirology 55 (2) (2012) 98–101.
[45] Anis Sharafoddini, Joel A. Dubin, David M. Maslove, Joon Lee, et al., A new insight into missing data in intensive care unit patient profiles: Observational study, JMIR Med. Inf. 7 (1) (2019) e11605.
[46] Diane E. Bennett, Silvia Bertagnolio, Donald Sutherland, Charles F. Gilks, The World Health Organisation's global strategy for prevention and assessment of HIV drug resistance, Antivir. Ther. 13 (2_suppl) (2008) 1–13.
[47] World Health Organisation, Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: Recommendations for a public health approach, 2016, Access through https://www.who.int/publications/i/item/9789241549684.
[48] Mavuto M. Mukaka, A guide to appropriate use of correlation coefficient in medical research, Malawi Med. J. 24 (2012) 69–71.
[49] Max Welling, Herding dynamical weights to learn, in: The International Conference on Machine Learning, 2009, pp. 1121–1128.
[50] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, Christoph H. Lampert, Icarl: Incremental classifier and representation learning, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2001–2010.
[51] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[52] Alex Graves, Santiago Fernández, Jürgen Schmidhuber, Bidirectional LSTM networks for improved phoneme classification and recognition, in: The International Conference on Artificial Neural Networks, 2005, pp. 799–804.
[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[54] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, Jason Weston, Curriculum learning, in: The International Conference on Machine Learning, 2009, pp. 41–48.
[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 6000–6010.
[56] Jacob Devlin Ming-Wei Chang Kenton, Lee Kristina Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: The Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.
[57] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, Improving language understanding by generative pre-training, in: A Technical Report of OpenAI, 2018.
[58] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, Ole Winther, Autoencoding beyond pixels using a learned similarity metric, in: The International Conference on Machine Learning, 2016, pp. 1558–1566.
[59] Mi-Ja Woo, Jerome P. Reiter, Anna Oganian, Alan F. Karr, Global measures of data utility for microdata masked for disclosure limitation, J. Priv. Confid. 1 (2009).
[60] Richard A. Davis, Keh-Shin Lii, Dimitris N. Politis, Remarks on some nonparametric estimates of a density function, in: Selected Works of Murray Rosenblatt, 2011, pp. 95–100.
[61] John L. Hodges, The significance probability of the Smirnov two-sample test, Arkiv För Matematik 3 (5) (1958) 469–486.
[62] Karen K. Yuen, The two-sample trimmed t for unequal population variances, Biometrika 61 (1974) 165–170.
[63] George W. Snedecor, Witiiam G. Cochran, Statistical methods, Ames: Iowa State Univ. Press Iowa 54 (1989) 71–82.
[64] Friedrich Pukelsheim, The three sigma rule, Amer. Statist. 48 (1994) 88–91.
[65] Maurice G. Kendall, The treatment of ties in ranking problems, Biometrika 33 (1945) 239–251.
[66] Scott Fujimoto, David Meger, Doina Precup, Off-policy deep reinforcement learning without exploration, in: The International Conference on Machine Learning, 2019, pp. 2052–2062.
[67] Sonali Parbhoo, Jasmina Bogojeska, Maurizio Zazzi, Volker Roth, Finale Doshi-Velez, Combining kernel and model based learning for HIV therapy selection, in: AMIA Summits on Translational Science Proceedings, vol. 2017, 2017, p. 239.
[68] Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D Mooney, Bradley A. Malin, A multifaceted benchmarking of synthetic electronic health record generation models, Nature Commun. 13 (1) (2022) 7609.
[69] Health Canada, Guidance document on public release of clinical information, 2014, Access through https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html.
[70] European Medicines Agency, European medicines agency policy on publication of clinical data for medical products for human use, 2014, Access through http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.
[71] Nicole M. Thomasian, Carsten Eickhoff, Eli Y. Adashi, Advancing health equity with artificial intelligence, J. Public Health Policy 42 (2021) 602–611.
[72] Karan Bhanot, Miao Qi, John S. Erickson, Isabelle Guyon, Kristin P. Bennett, The problem of fairness in synthetic healthcare data, Entropy 23 (9) (2021) 1165.
[73] Australian Government Department of Industry, Science, and Resources, Australia's artificial intelligence ethics framework, 2019, Access through https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework.
[74] United States of America Government Food and Drug Administration, Artificial intelligence/machine learning (AI/ML)-Based software as a medical device (SaMD) action plan, 2021, Access through https://www.fda.gov/media/145022/download.
[75] Sergey Levine, Aviral Kumar, George Tucker, Justin Fu, Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020, Preprint at https://arxiv.org/abs/2005.01643.
[76] Mike Wu, Sonali Parbhoo, Michael C Hughes, Volker Roth, Finale Doshi-Velez, Optimizing for interpretability in deep neural networks with tree regularization, J. Artificial Intelligence Res. 72 (2021) 1–37.
[77] Prafulla Dhariwal, Alexander Nichol, Diffusion models beat GANs on image synthesis, in: The Advances in Neural Information Processing Systems, 2021, pp. 8780–8794.
[78] Nicholas I. Kuo, Louisa Jorm, Sebastiano Barbieri, Synthetic health-related longitudinal data with mixed-type variables generated using diffusion models, 2023, Access through https://arxiv.org/abs/2303.12281.
[79] Nicholas I. Kuo, The health gym v2.0 synthetic antiretroviral therapy (ART) for HIV dataset, 2023, Access through https://figshare.com/articles/dataset/The_Health_Gym_v2_0_Synthetic_Antiretroviral_Therapy_ART_for_HIV_Dataset/22827878.
[80] Mattia C.F. Prosperi, Michal Rosen-Zvi, André Altmann, Maurizio Zazzi, Simona Di Giambenedetto, Rolf Kaiser, Eugen Schülter, Daniel Struck, Peter Sloot, David A Van De Vijver, et al., Antiretroviral therapy optimisation without genotype resistance testing: A perspective on treatment history based models, PLoS One 5 (2010) e13753.