Synthetic Health-related Longitudinal Data with Mixed-type Variables Generated using Diffusion Models

Nicholas I-Hsien Kuo[1], **Federico Garcia**[2a,2b,2c], **Anders Sönnerborg**[3], **Michael Böhm**[4], **Rolf Kaiser**[4], **Maurizio Zazzi**[5], **EuResist Network study group**, **Louisa Jorm**[1], **Sebastiano Barbieri**[1]

[1]Centre for Big Data Research in Health, the University of New South Wales, Sydney, Australia
[2a]Instituto de Investigación Ibs.Granada, Spain
[2b]Hospital Universitario San Cecilio, Spain
[2c]CIBER de Enfermedades Infecciosas, Spain
[3]Hospital Karolinska Institutet, Sweden
[4]Uniklinik Köln, Universität zu Köln, Germany
[5]Università degli Studi di Siena, Italy

Corresponding author: Nicholas I-Hsien Kuo (n.kuo@unsw.edu.au)

Abstract

This paper introduces a novel method for simulating Electronic Health Records (EHRs) using Diffusion Probabilistic Models (DPMs). We showcase the ability of DPMs to generate longitudinal EHRs with mixed-type variables – numeric, binary, and categorical. Our approach is benchmarked against existing Generative Adversarial Network (GAN)-based methods in two clinical scenarios: management of acute hypotension in the intensive care unit and antiretroviral therapy for people with human immunodeficiency virus. Our DPM-simulated datasets not only minimise patient disclosure risk but also outperform GAN-generated datasets in terms of realism. These datasets also prove effective for training downstream machine learning algorithms, including reinforcement learning and Cox proportional hazards models for survival analysis.[1]

1 Introduction

Machine learning (ML) plays a key role in realising personalised healthcare [1], but ML research and development are often hindered by privacy regulations limiting access to real-world datasets [2; 3]. Generative Adversarial Networks (GANs) [4] offer a solution by creating synthetic healthcare datasets [5]. However, GANs suffer from unstable training and mode collapse, which reduce data quality and diversity [6; 7]. These issues persist despite several mitigation techniques [8; 9; 10; 11] and can introduce biases to simulated data [12; 13; 14] that pose a potential risk of patient harm in downstream healthcare ML applications [15].

Recently, Diffusion Probabilistic Models (DPMs) [16; 17] have emerged as a promising alternative to GANs, offering better data realism in image synthesis [18; 19]. Unlike GANs, DPMs are not known to experience issues of unstable training and mode collapse. Early applications of DPMs in healthcare have successfully imputed tabular electronic health records (EHRs) [20].

This study aims to extend the scope of DPM applications to the generation of synthetic longitudinal EHRs with mixed-type variables. We demonstrate that our DPM-based approach minimises patient disclosure risks, enhances data realism, and proves effective for training downstream ML algorithms. Specifically, we show the utility of these synthetic datasets in Reinforcement Learning (RL) [21] to

[1]Refer to Section 6 for the ethics approval, broader impact, data access, and code repository of this paper.

NeurIPS 2023 Workshop on Synthetic Data for Empowering ML Research.

inform patient medication management and in Cox Proportional Hazards (CPH) models for survival analysis [22]. We further demonstrate that our synthetic healthcare datasets realistically represent important clinical milestones defined by the World Health Organisation (WHO) guidelines [23]. Our synthetic data thus can contribute to accelerating innovation in healthcare by facilitating development and benchmarking of ML algorithms, and supporting medical education [24].

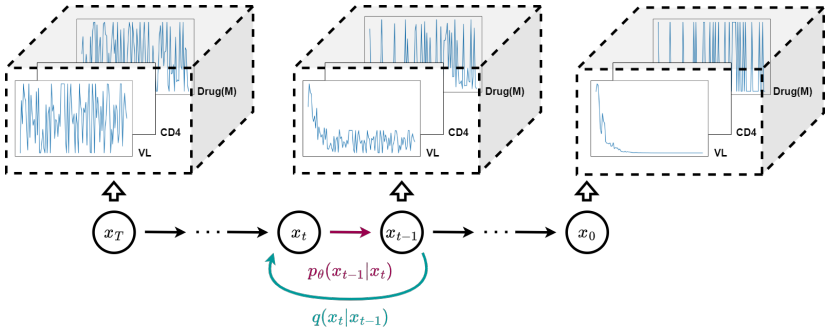


Figure 1: Using the DPM framework to generate synthetic sequential data.

2 Background

2.1 Diffusion Probabilistic Models

DPMs approximate real data distributions using diffusion and denoising, as shown in Figure 1. The forward diffusion process iteratively adds Gaussian noise

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{(1-\beta_t)}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

to a sample across T time-steps, following a pre-defined variance schedule $\{\beta_t \in (0,1)\}_{t=1}^T$. The iterative denoising process learns a model p_θ to approximate the conditional probability of the real data given the noisy input:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \;\text{ and }\; p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)). \tag{2}$$

Perturbed inputs with the added noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are written as

$$x_t(x_0, \epsilon) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Ho *et al.* [17] showed that the forward process is tractable when conditioned on x_0, and that a DPM could configure $\mu_\theta(x_t, t)$ to predict the noise in x_t with

$$\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}\right). \tag{4}$$

Furthermore in [17], they demonstrated that optimising the loss

$$\mathcal{L}_{\text{Noise}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\|^2\right] \tag{5}$$

is equivalent to optimising the negative log-likelihood using the variational lower bound.

To generate novel data, we follow Song *et al.* [25]'s score-based generative method using Langevin dynamics and sampling $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ where

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}\right) + \sigma_t\mathbf{z}. \tag{6}$$

2.2 Ground Truth Datasets

We generated a synthetic dataset for people receiving antiretroviral therapy for human immunodeficiency virus (ART for HIV) using EuResist [26] with published inclusion/exclusion criteria [5; 27]. The original dataset comprises 8,916 individuals, post-2015, on the 50 most common drug combinations. It includes demographics, viral load, CD4 counts, and treatment regimens. Data were rounded to the nearest 10-month interval, ranging from 10 to 100 months. Data missingness is high in the real dataset, and hence we included binary variables with suffix (M)s to denote data measurements at specific time-points. See § A.1 in the Supplementary Materials for more details.

In addition, we present findings on an acute hypotension dataset derived from MIMIC-III [28]. For details refer to § A.2 in the Supplementary Materials.

3 Methods

3.1 Backbone

Following prior studies [29; 30; 31], we adopt U-Net [32] as the backbone for generating mixed-type longitudinal time series data. Our U-Net utilises multi-layered 1D convolutional neural networks and autoencodes clinical time-series data in a bidirectional manner which is inspired by BERT's role in natural language processing [33]. More details are in the Supplementary Materials: § B for data transformation, § C for U-Net module selection, and § D for hyperparameter choices.

3.2 Auxiliary Loss Functions

Simulating medical time-series data offers distinct challenges due to sparsity and negative correlations [34]. To address the slow sampling of DPMs [35], we propose two auxiliary loss functions.

The first, a one-step reconstruction loss $\mathcal{L}_{\text{Recon}_1}$, is defined as

$$\mathcal{L}_{\text{Recon}_1} = \|x_0 - \hat{x}_0(t, \epsilon_\theta)\|_2^2, \text{ where } \hat{x}_0(t, \epsilon_\theta) = \frac{x_t - \sqrt{(1-\bar{\alpha}_t)}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}. \tag{7}$$

This leverages predicted noise ϵ_θ to approximate x_0, reducing computational overhead [25; 36].

The second, a latent discriminative projection loss $\mathcal{L}_{\text{Recon}_2}$, is

$$\mathcal{L}_{\text{Recon}_2} = \|\mathfrak{U}(x_0) - \mathfrak{U}(\hat{x}_0(t, \epsilon_\theta))\|_2^2, \text{ where } \mathfrak{U}(v) = \max(0, v\,\mathcal{U}_1)\,\mathcal{U}_2. \tag{8}$$

Inspired by mini-batch discrimination [11], this minimises x_0 and $\hat{x}_0$ discrepancy in a random feature space using untrained matrices $\mathcal{U}_1$ and $\mathcal{U}_2$, thereby reducing latent variability.

3.3 Metrics

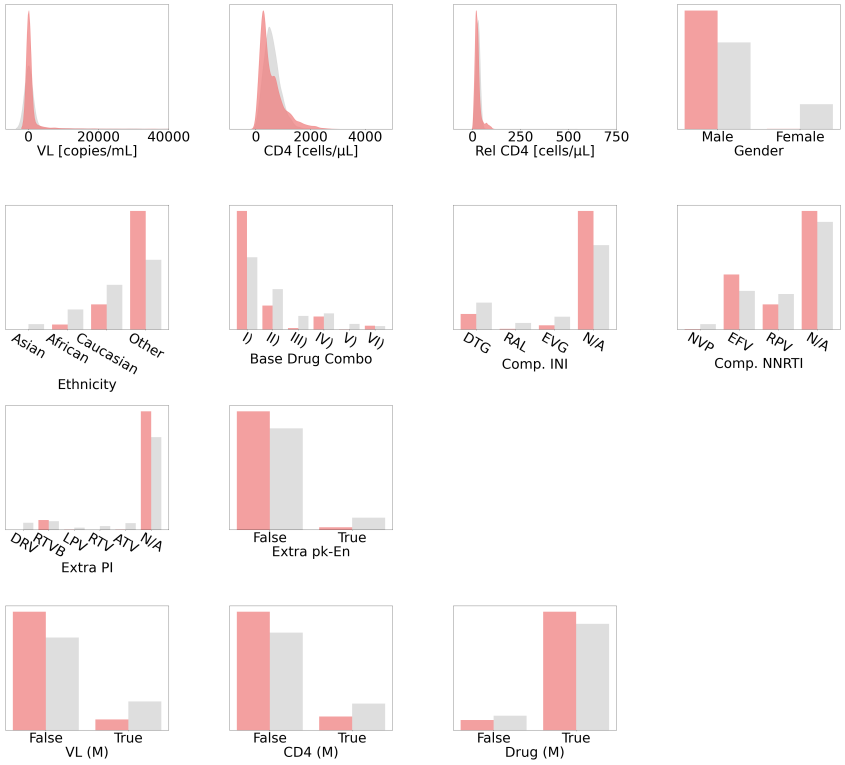
For variable fidelity, we utilise Kernel Density Estimations (KDEs) [37] for numerical variables and side-by-side barplots for binary and categorical variables. We confirm realism via the Kolmogorov-Smirnov (KS) test [38], Student's t-test [39], F-test [40], and three sigma rule tests [41]. Inter-variable interactions are inspected using Kendall's τ correlation [42].

Diversity is checked using log-cluster U [13] and category coverage (CAT) [14]: lower U and higher CAT are preferable. Notably, vision-specific metrics like Inception Score (IS) [11] and Fréchet Inception Distance (FID) [43] are not suitable for synthetic EHR evaluation. Data leakage is checked via minimum Euclidean distance [5]; privacy compliance is confirmed through sample-to-population attack [44], adhering to European [45] and Canadian [46] standards requiring risk below 9%.

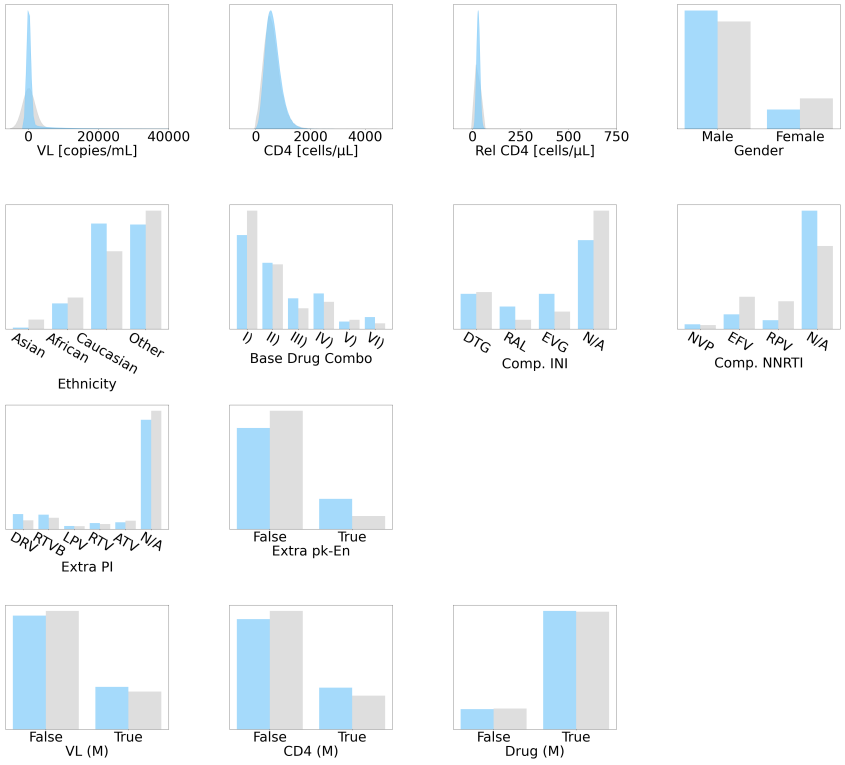
Utility evaluation involves comparing RL agents [47] and CPH models [48] trained on both real and synthetic datasets. For details in validation and utility setups, see § E in the Supplementary Materials.

4 Results

We compared our approach with the WGAN-GP model [9; 10] implemented in the Health Gym GAN [5; 49]. For clarity, we define: $\mathfrak{D}_{\text{real}}$ for the ground truth dataset, $\mathfrak{D}_{\text{null}}$ for the synthetic dataset generated via WGAN-GP [5], and $\mathfrak{D}_{\text{alt}}$ for our alternative synthetic dataset simulated using DPM.



(a) Synthetic dataset $\mathfrak{D}_{\text{null}}$ from [5]'s WGAN-GP in pink.



(b) Synthetic dataset $\mathfrak{D}_{\text{alt}}$ from our DPM in blue.

Figure 2: Comparing the variables in ART for HIV, with those of $\mathfrak{D}_{\text{real}}$ in colour grey.

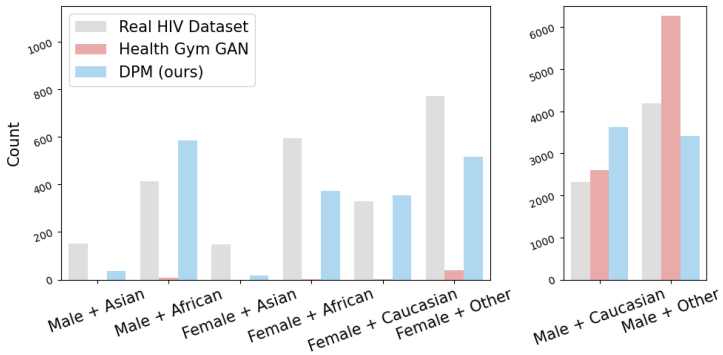


Figure 3: Comparing the patient demographics in the ART for HIV datasets.

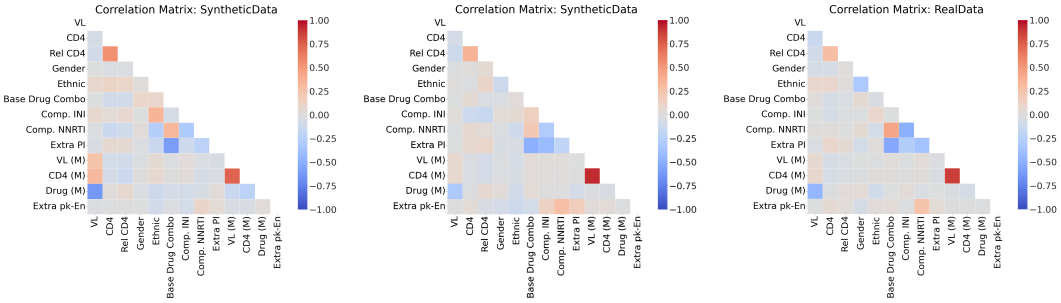
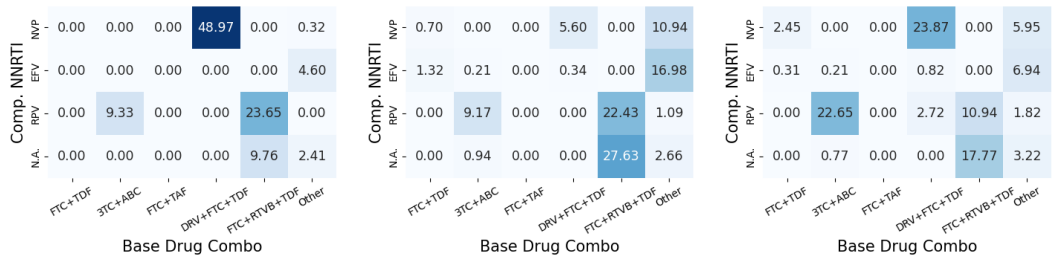
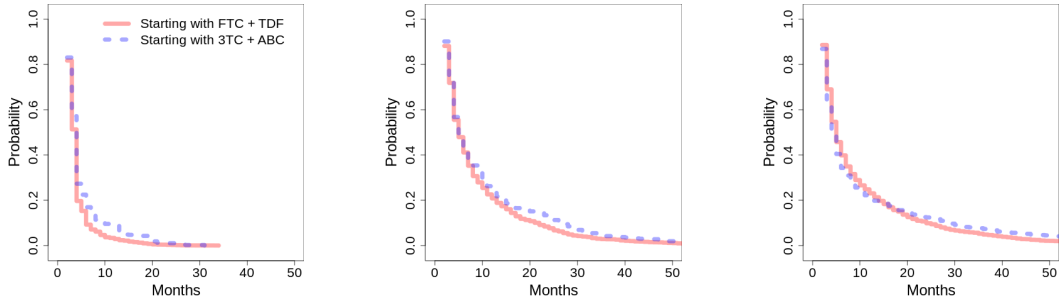


Figure 4: Comparing the correlations in ART for HIV in $\mathfrak{D}_{\text{null}}$ (left), $\mathfrak{D}_{\text{alt}}$ (middle), and $\mathfrak{D}_{\text{real}}$ (right).



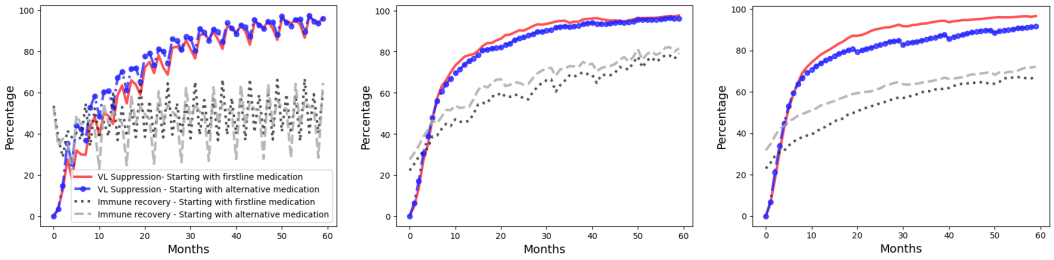
(a) With GAN-generated $\mathfrak{D}_{\text{null}}$. (b) With our DPM-simulated $\mathfrak{D}_{\text{alt}}$. (c) With the real dataset $\mathfrak{D}_{\text{real}}$.

Figure 5: Comparing the policies learned by RL agents using different ART for HIV datasets.



(a) On GAN-generated $\mathfrak{D}_{\text{null}}$. (b) On our DPM-simulated $\mathfrak{D}_{\text{alt}}$. (c) On the real dataset $\mathfrak{D}_{\text{real}}$.

Figure 6: Estimation of viral control via CPHs modelled on different ART for HIV datasets.



(a) In GAN-generated $\mathfrak{D}_{\text{null}}$. (b) In our DPM-simulated $\mathfrak{D}_{\text{alt}}$. (c) In the real dataset $\mathfrak{D}_{\text{real}}$.

Figure 7: Statistics for viral suppression and immune recovery on different ART for HIV datasets.

Dataset Realism
Figure 2 shows KDE plots[2] and barplots for ART for HIV. Our DPM-simulated $\mathfrak{D}_{\text{alt}}$ better approximates $\mathfrak{D}_{\text{real}}$, notably in imbalanced `Base Drug Combo` labels. All variables in $\mathfrak{D}_{\text{alt}}$ passed the KS test except `VL`, which nonetheless passed the three sigma rule test hence all variables in $\mathfrak{D}_{\text{alt}}$ achieved a high level of realism and reliability. For all outputs of the statistical tests, refer to § F.1 in the Supplementary Materials.

	U (↓)	**CAT** (↑)
$\mathfrak{D}_{\text{null}}$ [5]	-2.130	97.50%
$\mathfrak{D}_{\text{alt}}$ (ours)	**-3.057**	**100.00%**

Table 1: Metric comparison.

The statistics in Table 6 confirm $\mathfrak{D}_{\text{alt}}$ outperforms $\mathfrak{D}_{\text{null}}$ in lower U and higher CAT values, better capturing $\mathfrak{D}_{\text{real}}$'s latent structure. These metrics are contextualised in Figure 3, showing that $\mathfrak{D}_{\text{alt}}$ covers all demographic feature combinations, unlike $\mathfrak{D}_{\text{null}}$, suggesting its superior data heterogeneity. In addition, Figure 4 shows that both synthetic datasets mirror the correlations in the real dataset.

Private Information Disclosure Risk
Using Euclidean distance tests, we found no data leakage in $\mathfrak{D}_{\text{alt}}$, with a minimum distance of 0.09 ($>$0). Moreover, the sample-to-population attack risk was 0.076%, well below the 9% threshold [45; 46], ensuring safe distribution of our synthetic dataset.

Utility
Figure 5 visualises RL agents' learned policy for selecting combinations of `Comp. NNRTI` and `Base Drug Combo` for ART. The numbers on the heatmap represent the frequency of taking a specific action as a proportion of all actions taken. The RL agent trained on $\mathfrak{D}_{\text{null}}$ suggested (NVP, DRV + FTC + TDF) for 48.97% of all actions. This suggests that the GAN model used to generate $\mathfrak{D}_{\text{null}}$ experienced mode collapse, thus creating an excessive number of synthetic records with similar attributes in $\mathfrak{D}_{\text{null}}$. Conversely, we attribute the higher utility of the DPM-simulated $\mathfrak{D}_{\text{alt}}$ to the higher robustness of DPM against mode collapse.

We further focused on the effectiveness of ART combinations in controlling viral load, stratifying by different ART types (*i.e.,* `FTC + TDF` vs `3TC + ABC`) at initiation and the time needed to achieve a viral load under 1,000 copies/mL – a significant clinical milestone [23]. Figure 6 reveals CPH models built using $\mathfrak{D}_{\text{alt}}$ better emulate those on $\mathfrak{D}_{\text{real}}$ for predicting the required time. Models built using $\mathfrak{D}_{\text{null}}$ skewed towards unlikely early `VL` control.

In addition, we examined patient recovery trajectories over a 60-month span. Patients are categorised based on whether they initiated treatment with WHO-recommended first-line medications or alternative options [23]. We then visualised monthly percentages of patients achieving viral suppression (`VL` $<$ 200 copies/mL) and immune recovery (`CD4` $\geq$ 500 cells/μL) in each category, providing an initial baseline for comparison. Figure 7 shows that our DPM-simulated $\mathfrak{D}_{\text{alt}}$ closely mirrors the real dataset $\mathfrak{D}_{\text{real}}$, while GAN-generated $\mathfrak{D}_{\text{null}}$ exhibits unexplained strong seasonality.

Details for the experimental setups are in § E.4 of the Supplementary Materials. Refer to more results comparing synthetic datasets for acute hypotension in § F.2 of the Supplementary Materials.

[2]The kernel density estimation uses Gaussian kernels to estimate the probability density function of a continuous variable. Thus, the KDE function can potentially produce tails beyond the range of the data.

5 Discussion

This paper introduced a DPM framework featuring a custom U-Net backbone and two auxiliary loss functions to generate mixed-type longitudinal clinical datasets. We demonstrated its superior utility and realism compared to GAN-generated datasets, while minimising patient disclosure risk, thereby facilitating open access to reliable healthcare data for research and medical education.

Although existing synthetic data studies predominantly focus on image-related tasks, clinical data, with its unique challenges, remains underexplored. Most of the existing studies either synthesise a single data type or produce static datasets [12; 50; 51; 52; 53; 54; 55; 56]. While some studies have produced longitudinal EHRs [57; 5], they neglect in-depth analysis of their datasets' time-dependent nature. Our contribution lies in demonstrating that our DPM-simulated clinical datasets not only uphold realism and minimise private information disclosure risk, but also offer high utility. This is evidenced by their performance in downstream RL agents (Figure 5) and CPH models (Figure 6), as well as their underlying characteristics when inspected using WHO guidelines (Figure 7).

6 Ethics Approval, Broader Impact, Data Access, and Code Repository

We applied DPMs to longitudinal data extracted from the MIMIC-III [28] and the EuResist [26] databases to generate our synthetic datasets. This study was approved by the University of New South Wales' human research ethics committee (application HC210661). For patients in MIMIC-III requirement for individual consent was waived because the project did not impact clinical care and all protected health information was deidentified [28]. For people in the EuResist integrated database all data providers obtained informed consent for the execution of retrospective studies and inclusion in merged cohorts [58].

The broader adoption of synthetic data in medical education and research represents a paradigm shift that can profoundly transform the landscape of health data science [24]. By offering controlled, context-specific resources that closely emulate real-world scenarios, synthetic data provides an invaluable resource for students and researchers, allowing them to gain practical experience without risking patient confidentiality. Such datasets democratise access to vital information, breaking barriers imposed by stringent privacy regulations, and equipping future health professionals with the tools they need to drive innovation in healthcare AI and analytics. However, while these datasets offer promise, we emphasise that synthetic datasets should not be naïvely used to replace real datasets; and that critical evaluation in diverse applications will be vital to determine their ultimate efficacy in clinical research.

Our superior synthetic datasets generated using DPMs that we introduced in this paper are now available on **Health Gym** [5], accessible at `https://healthgym.ai/`, the same platform that hosted the baseline synthetic datasets generated using GAN-based methods. The project's open-source code is also available on GitHub at
`https://github.com/Nic5472K/ScientificData2021_HealthGym`.

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).

References

[1] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature medicine*, vol. 24, pp. 1716–1720, 2018.

[2] R. Nosowsky and T. J. Giordano, "The health insurance portability and accountability act of 1996 (hipaa) privacy rule: Implications for clinical research," *Annual Review of Medicine*, vol. 57, pp. 575–590, 2006.

[3] C. M. O'Keefe and C. J. Connolly, "Privacy and the use of health data for research," *Medical Journal of Australia*, vol. 193, no. 9, pp. 537–541, 2010.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *the Advances in Neural Information Processing Systems*, 2014.

[5] N. I. Kuo, M. N. Polizzotto, S. Finfer, F. Garcia, A. Sönnerborg, M. Zazzi, M. Böhm, R. Kaiser, L. Jorm, S. Barbieri *et al.*, "The health gym: Synthetic health-related datasets for the development of reinforcement learning algorithms," *Scientific Data*, vol. 9, no. 1, pp. 1–24, 2022.

[6] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of gans," 2017, preprint at https://arxiv.org/abs/1705.07215.

[7] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *the International Conference on Machine Learning*, 2018.

[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, preprint at https://arxiv.org/abs/1511.06434.

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *the International Conference on Machine Learning*, 2017.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *the Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017.

[11] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *the Advances in Neural Information Processing Systems*, 2016.

[12] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *the Machine Learning for Healthcare Conference*, 2017.

[13] M.-J. Woo, J. P. Reiter, A. Oganian, and A. F. Karr, "Global measures of data utility for microdata masked for disclosure limitation," *Journal of Privacy and Confidentiality*, vol. 1, 2009.

[14] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Medical Research Methodology*, vol. 20, pp. 1–40, 2020.

[15] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias, and clinical safety," *BMJ Quality & Safety*, vol. 28, pp. 231–237, 2019.

[16] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *the International Conference on Machine Learning*, 2015.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *the Advances in Neural Information Processing Systems*, 2020.

[18] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022, preprint at https://arxiv.org/abs/2204.06125.

[19] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *the Advances in Neural Information Processing Systems*, 2021.

[20] S. Zheng and N. Charoenphakdee, "Diffusion models for missing value imputation in tabular data," in *the First Table Representation Workshop of the Advances in Neural Information Processing Systems*, 2022.

[21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[22] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, 1972.

[23] 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.

[24] N. I. Kuo, O. Perez-Concha, M. Hanly, E. Mnatzaganian, B. Hao, M. Di Sipio, G. Yu, J. Vanjara, I. C. Valerie, J. De Oliveira Costa, T. Churches, S. Lujic, J. Hegarty, L. Jorm, and S. Barbieri, "Enriching data science and healthcare education: Application and impact of synthetic datasets through the health gym project," 2023, access through https://doi.org/10.2196/preprints.51388.

[25] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *the Advances in Neural Information Processing Systems*, 2019.

[26] M. Zazzi, F. Incardona, M. Rosen-Zvi, M. Prosperi, T. Lengauer, A. Altmann, A. Sonnerborg, T. Lavee, E. Schülter, and R. Kaiser, "Predicting response to antiretroviral treatment by machine learning: The euresist project," *Intervirology*, vol. 55, no. 2, pp. 123–127, 2012.

[27] S. Parbhoo, J. Bogojeska, M. Zazzi, V. Roth, and F. Doshi-Velez, "Combining kernel and model based learning for hiv therapy selection," *AMIA Joint Summits on Translational Science proceedings*, vol. 2017, p. 239, 2017.

[28] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, pp. 1–9, 2016.

[29] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *the ACM Special Interest Group on Computer Graphics*, 2022.

[30] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *the ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[31] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. 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.

[34] N. I. Kuo, L. Jorm, S. Barbieri *et al.*, "Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for hiv," 2022, preprint at https://arxiv.org/abs/2208.08655.

[35] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," in *the International Conference on Learning Representations*, 2022.

[36] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *the International Conference on Learning Representations*, 2021.

[37] R. A. Davis, K.-S. Lii, and D. N. Politis, "Remarks on some nonparametric estimates of a density function," in *Selected Works of Murray Rosenblatt*, 2011, pp. 95–100.

[38] J. L. Hodges, "The significance probability of the smirnov two-sample test," *Arkiv för Matematik*, vol. 3, no. 5, pp. 469–486, 1958.

[39] K. K. Yuen, "The two-sample trimmed t for unequal population variances," *Biometrika*, vol. 61, pp. 165–170, 1974.

[40] G. W. Snedecor and W. G. Cochran, "Statistical methods," *Ames: Iowa State University Press*, vol. 54, pp. 71–82, 1989.

[41] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, pp. 88–91, 1994.

[42] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, pp. 239–251, 1945.

[43] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *the Advances in Neural Information Processing Systems*, 2017.

[44] K. El Emam, L. Mosquera, and J. Bass, "Evaluating identity disclosure risk in fully synthetic health data: Model development and validation," *Journal of Medical Internet Research*, vol. 22, p. 23139, 2020.

[45] 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 .

[46] 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.

[47] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *the International Conference on Machine Learning*, 2019.

[48] T. M. Therneau and T. Lumley, "Package 'survival'," *R Top Doc*, vol. 128, no. 10, 2015.

[49] N. Kuo, "The Heath Gym Synthetic HIV Dataset," 7 2023. [Online]. Available: https://figshare.com/articles/dataset/The_Heath_Gym_Synthetic_HIV_Dataset/19838470

[50] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," 2018, preprint at https://arxiv.org/abs/1802.06739.

[51] R. Camino, C. Hammerschmidt, and R. State, "Generating multi-categorical samples with generative adversarial networks," in *the ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

[52] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 7, p. e005122, 2019.

[53] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proceedings of the VLDB Endowment*, vol. 11, no. 10, pp. 1071–1083, 2018.

[54] P.-H. Lu, P.-C. Wang, and C.-M. Yu, "Empirical evaluation on synthetic data generation with generative adversarial network," in *the Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, 2019.

[55] J. Yoon, L. N. Drumright, and M. Van Der Schaar, "Anonymization through data synthesis using generative adversarial networks (ads-gan)," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2378–2388, 2020.

[56] M. Walia, B. Tierney, and S. McKeever, "Synthesising tabular data using wasserstein conditional gans with gradient penalty (wcgan-gp)." in *AICS*, 2020.

[57] J. Li, B. J. Cairns, J. Li, and T. Zhu, "Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications," 2021, preprint at https://arxiv.org/abs/2112.12047.

[58] M. C. Prosperi, M. Rosen-Zvi, A. Altmann, M. Zazzi, S. Di Giambenedetto, R. Kaiser, E. Schülter, D. Struck, P. Sloot, D. A. Van De Vijver *et al.*, "Antiretroviral therapy optimisation without genotype resistance testing: A perspective on treatment history based models," *PloS one*, vol. 5, p. e13753, 2010.

[59] O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. A. Celi, "Guidelines for reinforcement learning in healthcare," *Nature Medicine*, vol. 25, pp. 16–18, 2019.

[60] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, pp. 259–284, 1998.

[61] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, preprint at https://arxiv.org/abs/1301.3781.

[62] A. Mottini, A. Lheritier, and R. Acuna-Agost, "Airline passenger name record generation using generative adversarial networks," 2018, preprint at https://arxiv.org/abs/1807.06657.

[63] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *the International Conference on Machine Learning*, 2009.

[64] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *the European Conference on Computer Vision*, 2014.

[65] W. N. Venables and B. D. Ripley, *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *the Advances in Neural Information Processing Systems*, 2017.

[67] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding with Unsupervised Learning," *Technical Report, OpenAI*, 2018.

[68] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, preprint at https://arxiv.org/abs/1607.06450.

[69] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[70] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, preprint at https://arxiv.org/abs/1312.4400.

[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, preprint at https://arxiv.org/abs/1412.6980.

[72] M. Hernadez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions," *Methods of Information in Medicine*, 2023.

[73] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[74] R. Liu, J. L. Greenstein, J. C. Fackler, J. Bergmann, M. M. Bembea, and R. L. Winslow, "Offline reinforcement learning with uncertainty for treatment strategies in sepsis," 2021, preprint at https://arxiv.org/abs/2107.04491.

[75] J. A. Wegelin, "A survey of partial least squares (pls) methods, with emphasis on the two-block case," University of Washington, Tech. Rep., 2000.

[76] S. Vassilvitskii and D. Arthur, "k-means++: The ddvantages of careful seeding," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2006.

Supplementary Materials for the paper
Synthetic Health-related Longitudinal Data with Mixed-type Variables Generated using Diffusion Models

Anonymous Author(s)

In the main paper, we introduced a novel approach for simulating electronic health records (EHRs) using diffusion probabilistic models (DPMs) [16; 17], demonstrating their effectiveness in generating longitudinal EHRs with mixed-type variables including numeric, binary, and categorical variables. We compared the performance of our DPM-simulated datasets with state-of-the-art generative adversarial networks (GANs) [4] for two clinical applications: management of patients with acute hypotension in the intensive care unit (ICU) and antiretroviral therapy (ART) for human immunodeficiency virus (HIV). Moreover, we trained reinforcement learning (RL) agents [21] and Cox proportional hazard (CPH) models [22] on the synthetic data to evaluate the utility of our approach for developing downstream machine learning models.

Due to the constraints on the length of the main manuscript, we have moved the majority of our implementation details and supplementary findings to this accompanying document.

Content of Tables:

A The Ground Truth Datasets

We based our work on the Health Gym project [5], which used GANs to generate synthetic longitudinal data from two health-related databases: MIMIC-III [28] and EuResist [26]. The authors used these databases to generate synthetic datasets for the management of acute hypotension and ART for HIV. The patient cohorts were defined using inclusion and exclusion criteria from previous studies: Gottesman *et al.* [59] for acute hypotension and Parbhoo *et al.* [27] for ART for HIV.

For data extraction and the inclusion/exclusion criteria, we mainly followed the Supplementary Information provided by [5] in `https://www.nature.com/articles/s41597-022-01784-7`. Additional guidelines on data formatting can be found in [5]'s repository `https://github.com/Nic5472K/ScientificData2021_HealthGym`.

A.1 ART for HIV

The real HIV dataset is based on a cohort of individuals from the EuResist database, as proposed by Parbhoo *et al.* [27]. The study employs a mixture-of-experts approach for therapy selection, utilising kernel-based methods to identify clusters of similar individuals and an RL agent to optimise treatment strategy. The dataset consists of 8,916 individuals who started therapy after 2015 and were treated with the 50 most common medication combinations, including 21 different types of medications. Demographics, viral load (VL), CD4 counts, and regimen information are included in the dataset. The length of therapy in the dataset varies, thus the records were truncated and modified to the closest multiples of 10-month periods, resulting in a shortest record length of 10 months and a longest record length of 100 months, each summarising patient observations over a 1-month time period. The dataset includes variables with suffix (M) to indicate the measurement at a specific point in time and is significant due to its informativeness in missing data in clinical time series, which can indicate the need for laboratory tests. Refer to Table 2 for more details.

Variable Name	**Data Type**	**Unit**	**Extra Notes**
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	- -	Female; Male
Ethnicity	categorical	- -	4 Classes Asian; African; Caucasian; Other
Base Drug Combination (Base Drug Combo)	categorical	- -	6 Classes I) FTC + TDF; II) 3TC + ABC; III) FTC + TAF; IV) DRV + FTC + TDF; V) FTC + RTVB + TDF; VI) Other
Complementary INI (Comp. INI)	categorical	- -	4 Classes DTG; RAL; EVG; Not Applied
Complementary NNRTI (Comp. NNRTI)	categorical	- -	4 Classes NVP; EFV; RPV; Not Applied
Extra PI	categorical	- -	6 Classes 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

Table 2: Variables in the ART for HIV Dataset.

A.2 Acute Hypotension

This dataset was extracted from MIMIC-III and was originally proposed by Gottesman *et al.* [59]. It comprises of 3,910 patients with 48-hour clinical variables, aggregated per hour in the time-series. The dataset includes variables with suffix (M) to indicate the measurement at a specific point in time and is significant due to its informativeness in missing data in clinical time series, which can indicate the need for laboratory tests. In their work, Gottesman *et al.* utilised this dataset to develop an RL agent that suggested optimal fluid boluses and vasopressors for acute hypotension management, with actions being made in a discrete action space by binning the boluses and vasopressors into multiple categories. Refer to Table 3 for more details.

Variable Name	Data Type	Unit	Extra Notes
Mean Arterial Pressure (MAP)	numeric	mmHg	
Diastolic Blood Pressure (DBP)	numeric	mmHg	
Systolic BP (SBP)	numeric	mmHg	
Urine	numeric	mL	
Alanine Aminotransferase (ALT)	numeric	IU/L	
Aspartate Aminotransferase (AST)	numeric	IU/L	
Partial Pressure of Oxygen (PaO_2)	numeric	mmHg	
Lactate	numeric	mmol/L	
Serum Creatinine	numeric	mg/dL	
Fluid Boluses	categorical	mL	4 Classes $[0, 250)$; $[250, 500)$; $[500, 1000)$; ≥ 1000
Vasopressors	categorical	mcg/kg/min	4 Classes 0; $(0, 8.4)$; $[8.4, 20.28)$; ≥ 20.28
Fraction of Inspired Oxygen (FiO_2)	categorical	fraction	10 Classes ≤ 0.2; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9; 1.0
Glasgow Coma Scale Score (GCS)	categorical	point	13 Classes 3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15
Urine Data Measured (Urine (M))	binary	- -	False; True
ALT or AST Data Measured (ALT/AST (M))	binary	- -	False; True
FiO_2 (M)	binary	- -	False; True
GCS (M)	binary	- -	False; True
PaO_2 (M)	binary	- -	False; True
Lactic Acid (M)	binary	- -	False; True
Serum Creatinine (M)	binary	- -	False; True

Table 3: Variables in the Acute Hypotension Dataset.

B Data Formulation for Mixed-Type Inputs & Outputs

For each iteration, we draw ground truth data ξ_0 from the set of clinical datasets and reformulate it to x_0 (to be addressed below). We also select a noise level t and its corresponding strength of perturbation β_t to introduce corruption to x_0 following Equation (1) in the main text to acquire the noisy inputs x_t. To estimate the manually injected noise ϵ of Equation (3) in the main text, we feed x_t into a tailored implementation of U-Net [32], which serves as our backbone network for the denoising operations. The output of the U-Net network ϵ_θ is the predicted estimation for ϵ.

Our datasets encompass numeric, binary, and categorical variables. Hence, we elaborate on the data formulation prior to presenting it to the model. The ground truth data is partitioned as $\xi_0 = \xi_{0,[\text{num}]} \oplus \xi_{0,[\text{alt}]}$, with the numeric subset $\xi_{0,[\text{num}]}$ and the non-numeric subset $\xi_{0,[\text{alt}]}$. We transform each numeric feature in $\xi_{0,[\text{num}]}$ to the range $\in [0, 1]$ and derive $x_{0,[\text{num}]}$. Each

non-numeric variable $\xi^{(i)}_{0,[\text{alt}]}$ is converted into a list of one-hot vectors where the observed class is assigned a value of 1 and the others a value of 0. Here are some examples:

For the binary variable `Gender = Female`, we have

$$\texttt{Gender} = \begin{bmatrix} \texttt{Female} \\ \texttt{Male} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ and}$$

for the categorical variable `Ethnicity = African`, we have

$$\texttt{Ethnicity} = \begin{bmatrix} \texttt{Asian} \\ \texttt{African} \\ \texttt{Caucasian} \\ \texttt{Other} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

We denote the aggregate of all one-hot vectors as $x_{0,[\text{alt}]} = \bigcup_i \text{OneHot}\left(\xi^{(i)}_{0,[\text{alt}]}\right)$; and that $x_0 = x_{0,[\text{num}]} \oplus x_{0,[\text{alt}]}$. Embeddings [60; 61; 62] are not necessary in our framework. The forward diffusion process of the DPM (refer to Equation (3) in the main text) directly applies noise to the one-hot vectors.

At test time, we randomly sample a noisy input x_T from a Gaussian distribution. Then, we iteratively estimate the corruption $\epsilon_\theta(x_t, t)$ at step t using our U-Net backbone to generate a less noisy x_{t-1} as per Equation (6) in the main text. Once we reach the allegedly clean and novel data x_0, we compartmentalise it into $x_{0,[\text{num}]}$ and $x_{0,[\text{alt}]}$ to reverse the transformation in $x_{0,[\text{num}]}$, resulting in $\xi_{0,[\text{num}]}$. Next, we employ softmax to recover the non-numeric variables such that $\xi_{0,[\text{alt}]} = \bigcup_i \text{Softmax}\left(x^{(i)}_{0,[\text{alt}]}\right)$.

The dimensionality of the noisy input is $x_t \in \mathbb{R}^{\mathfrak{B}\times 1\times \mathfrak{L}\times \mathfrak{N}}$, where $\mathfrak{B}$ corresponds to the batch size, $\mathfrak{L}$ denotes the length of the time-series, and that there is 1 feature channel for all the $\mathfrak{N}$ variables. As we previously mentioned, all acute hypotension data possess a fixed sequence with a length of 48 units, hence $\mathfrak{L}_{\text{hypotension}} = 48$. On the other hand, the HIV data has variable lengths and we utilise zero-padding to bring all the data to a pre-defined maximal length of $\mathfrak{L}_{\text{HIV}} = 100$. This setup hence obviates the need for curriculum learning [63] to enable training.

Moreover, inferring the size $\mathfrak{N}$ is a straightforward task, as it solely involves concatenating the numeric and one-hot representations of the binary and categorical variables in x_t. To illustrate, consider the ART for HIV dataset, whose variable specifications are provided in Table 2. By summing up the corresponding levels of every variable (with 1 for numeric variables), we obtain that
$\mathfrak{N}_{\text{HIV}} = \text{sum}(\{1, 1, 1, 2, 4, 6, 4, 4, 6, 2, 2, 2, 2\}) = 37$.
Likewise, we deduce that $\mathfrak{N}_{\text{hypotension}} = 54$ as per its respective specifications in Table 3.

C The U-Net Backbone

U-Net [32] is a convolutional neural network (CNN) architecture originally developed for medical image segmentation. As shown in Figure 8, the architecture has many details. The down-sampling [64] compartment extracts high-level features from noisy data, while skip connections [65] maintain fine-grained details and spatial information. The up-sampling compartment estimates noise for reconstructing clean data, leveraging localised features via the skip connections. U-Net is especially useful in denoising spatially correlated noise of varying intensities; and has been employed in various DPM applications [29; 30; 31]. We depicted the U-Net processing procedure in Figure 8.

C-a): Embedding the noise level
In Equation (4) in the main text, the noise prediction process of DPM is enabled via μ_θ to create ϵ_θ to predict noise ϵ. Notably, μ_θ is informed by noise level t, which is used to iteratively estimate noise across various levels. To this end, we adopt the Transformer sinusoidal position embedding method [66], as applied in Ho *et al.* [17], to featurise the noise level. These noise level embeddings are then incorporated into the U-Net architecture, and are fed as input to each intermediate neural activation stage that arises from the down- and up-sampling operations.

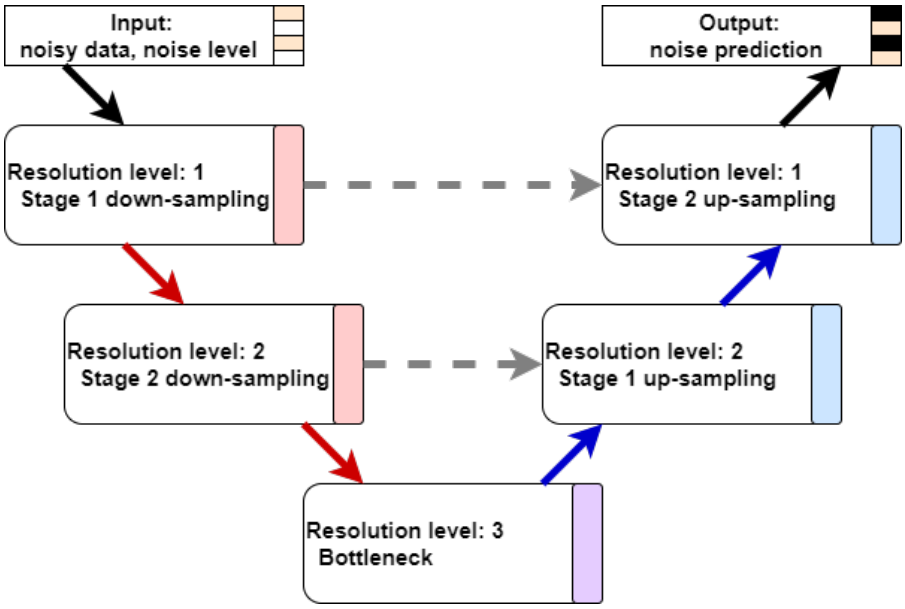


Figure 8: An overview of the elements of our U-Net.
Our U-Net is depicted in the top left panel, with the down-sampling, bottleneck, and up-sampling procedures denoted by the colors red, purple, and blue, respectively. Top right: The presence of linear transformations for pre- and post-processing. Bottom right: The local features in each resolution level is processed with block processing units and linear transformations.

C-b): Down- and up-sampling
All CNNs employed in our design are one-dimensional (1-D) and do not possess a causal architecture. Thus when we denoise the noisy acute hypotension datum $x_t \in \mathbb{R}^{\mathfrak{B}\times 1\times \mathfrak{L}\times \mathfrak{N}}$ with a fixed length of $\mathfrak{L} = 48$, the U-Net could simultaneously denoise the noisy data at positions 10 and 20. Our U-Net hence processes data similar to the autoencoding style of BERT [33], as opposed to the autoregressive style of GPT [67] (*i.e.,* we are not limited to denoising from left-to-right in a single direction). See more discussion in Section C-d).

C-c): Block feature extractor
After each stage of sampling operation, the noisy data is further processed while maintaining the same resolution level. Within each level, we utilise three successive feature extraction blocks, each composed of layer normalisation [68] followed by two 1-D CNNs.

C-d): Distinctive Additions to Our U-Net Architecture
We found that the application of the 1-D CNNs alone is insufficient for denoising. As elaborated upon in Section C-b), 1-D CNNs have the capability to denoise the noisy data simultaneously at positions 10 and 20, but for each feature independently. For ART for HIV, denoising VL is hence done independently of the regimen taken. While 2-D CNNs may seem more viable, an incorrect kernel size can still cause the erroneously denoising the of {VL, regimen} (in the kernel), while leaving out the relevant information of {CD4, Ethnicity} (out of the kernel). The need to concurrently denoise multiple time-series variables introduces a level of complexity that is not encountered in the DPM's application in speech [69].

This can be addressed by applying additional linear transformation layers on the $\mathfrak{N}$ dimension of x_t. As a consequence, the U-Net no longer denoises data at a variable level and instead denoises data on their latent features. Inspired by Lin *et al.* [70], we also include linear transformations to each up- and down-sampling 1-D CNN (see the bottom right panel of Figure 8) to process local patches within the receptive field.

Additional linear transformations are then employed on the final up-sampling output. This restructures the predicted noise made on the latent structure back to the $\mathfrak{N}$ sequences on x_t.

D Hyper-parameters

D.1 Hyper-parameters for the U-Net

Following Section C-d), we choose to linearly project the $\mathfrak{N}$ variables in the input to a latent space of dimensionality 256. After this preliminary step, we employ our U-Net for denoising.

As detailed in Section C-b), we adopt 3 distinct resolution levels. More specifically, resolution level 1 maintains the initial length of the noisy sequence for all time-series data, whereas the succeeding resolution levels condense the sequences while augmenting feature dimensions. In resolution level 2, all time-series have feature size 10, and in resolution level 3, all time-series possess feature size 20, regardless of the underlying dataset. However, the alteration in the length of the time-series depended on the dataset. For acute hypotension, resolution level 1 sequences span 48 time steps, which are subsequently reduced to 12 and 3 in resolution levels 2 and 3, respectively. Likewise in ART for HIV, they change from length 100 to 10 and then 3.

Following the previous descriptions, the 1-D CNNs employed in the blocks of Section C-c) possess feature dimensions of 10 and 20 at resolution levels 1 and 2 respectively. Whereas the features in the bottleneck of resolution level 3 reduces from 20 to 10, subsequently reverting to 20.

D.2 An Example using Acute Hypotension

The input data $x_t \in \mathbb{R}^{\mathfrak{B}\times 1\times 48\times 37}$ comprises a sequence length $\mathfrak{L}$ of 48 and $\mathfrak{N}$ variables of 37 (see Section B). We project and contruct the latent structure of the noisy data in $\mathbb{R}^{\mathfrak{B}\times 1\times 48\times 256}$. In the subsequent use of U-Net, the dimensionality transforms to $\mathbb{R}^{\mathfrak{B}\times 10\times 12\times 256}$ in resolution level 2 and then $\mathbb{R}^{\mathfrak{B}\times 20\times 3\times 256}$ in resolution level 3.

D.3 Hyper-parameters for the DPM & Optimisation

We set the maximum perturbation at $\beta_0 = 0.01$ and the minimum at $\beta_T = 1 \times 10^{-4}$ (see Equation (1) in the main text) across both datasets. However, we use $T = 1000$ for acute hypotension and $T = 500$ for ART for HIV. The intermediate perturbations β_t are distributed uniformly across the T levels. As previously stated in Section C-a), the denoising procedure of our DPM is informed by the noise level t. This information is conveyed to the U-Net architecture as a Transformer sinusoidal position embedding, featuring an embedding dimensionality of 100.

Our DPMs are updated using the Adam optimiser [71] with learning rate 1×10^{-3}. We employ a batch size of 128 for the acute hypotension and ART for HIV. The DPMs are trained for 5000 epochs for acute hypotension and 3000 epochs for ART for HIV. In addition, the losses are weighted at a ratio of $1 : 20 : 10$ for $\mathcal{L}_{\text{noise}}$, $\mathcal{L}_{\text{Recon}_1}$, and $\mathcal{L}_{\text{Recon}_2}$ (see Section 3.2 in the main text), respectively.

E Metrics

We put forth five desiderata:

Section E.1: that all generated variables to exhibit individual realism;
Section E.2: that there exists a sufficiently high level of diversity in variables;
Section E.3: that our synthetic datasets ensure patient privacy; and
Section E.4: that our datasets can function as a substitute for a genuine dataset in downstream model construction.

E.1 Assessing Individual Realisticness

We leverage two plots to assess the individual realisticness. For numeric variables, we use kernel density estimations (KDEs) [37] to overlay the synthetic distribution on top its genuine counterpart. For binary and categorical variables, we use side-by-side barplots to demonstrate the percentage share of each level.

Following Kuo *et al.* [5] and Hernadez *et al.* [72], we perform four statistical tests on the synthetic datasets shown in Figure 9. We begin with the two-sample Kolmogorov-Smirnov (KS) test [38] to evaluate whether the synthetic variables effectively capture the distributional characteristics of their

real counterparts. If a synthetic variable passes the KS test, it is deemed to be realistic and can be considered as having been drawn from the real datasets. Otherwise, we seek to identify the underlying reasons for its lack of realism.

The perceived lack of realism could be understood using the Student's t-test [39] and the F-test. Snedecor's F-test [40] is used for numeric variables; and we use the analysis of variance F-test for binary and categorical variables. The t-test verifies the alignment between means, while the F-test assesses the agreement in variances. However, in the event that a synthetic variable fails the KS test, neither the t-test nor the F-test can be used to assess the reliability of the synthetic variable. Hence, we choose the three sigma rule test [41] (by default, with 2 standard deviations) to evaluate whether the synthetic values fall within a plausible range of real values.

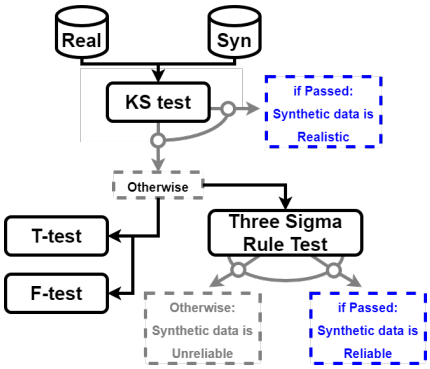


Figure 9: Statistical tests.
The sequence of the hypothesis tests.

Note, unlike image generation, we cannot employ the inception score (IS) [11] and the Fréchet inception distance (FID) [43] to evaluate the quality of our generated data. These metrics rely on the Inception v3 model [73], which is not suitable for analysing our longitudinal EHR data.

E.2 Evaluating Diversity on the Data Structure

To assess the level of diversity present in our synthetic datasets, we employ two metrics: the log-cluster metric U [13] and category coverage (CAT) proposed in Goncalves *et al.* [14]. The former, formulated as

$$U = \log\left(\frac{1}{\Gamma}\sum_{k=1}^{\Gamma}\left[\frac{n_{k_{\text{real}}}}{n_k} - \frac{n_{k_{\text{real}}}}{n_{k_{\text{real}}} + n_{k_{\text{syn}}}}\right]^2\right), \tag{9}$$

measures the difference in latent structures between the real and synthetic datasets. To compute U, we first sample records from both the real and synthetic datasets and then merge the sub-datasets to perform a cluster analysis via k-means with $\Gamma = 20$ clusters. Here, n_k represents the total number of records in cluster k, while $n_{k_{\text{real}}}$ and $n_{k_{\text{syn}}}$ denote the number of real and synthetic records in cluster k, respectively. We repeat this process 20 times for each synthetic dataset, with each repetition involving a sample of 100,000 real and synthetic records. A lower U score indicates that the synthetic datasets are more realistic.

The latter metric, CAT, is defined as

$$\text{CAT} = \frac{1}{J}\sum_{j=1}^{J}\frac{\|\mathfrak{D}_{\text{syn}}^{(j)}\|}{\|\mathfrak{D}_{\text{real}}^{(j)}\|}, \tag{10}$$

where J is the total number of binary and categorical variables, and $\mathfrak{D}_{\text{real}}^{(j)}$ and $\mathfrak{D}_{\text{syn}}^{(j)}$ represent the real and synthetic datasets, respectively, for the j-th variable. Specifically, CAT measures the completeness of the non-numeric classes in the synthetic datasets; it is the higher the better.

E.3 Security Estimation

We conduct two tests. First, we examine the minimum Euclidean distance between synthetic and actual records and verify that it is greater than zero, thus preventing any real records from being leaked into the synthetic dataset. Then, we utilise the sample-to-population attack in El Emam *et al.* [44] to assess the potential risk of an attacker learning new information by linking an individual in the synthetic dataset to the actual dataset.

The sample-to-population attack involves *quasi-identifiers*, which are variables that may reveal an individual's identity, such as `Gender` and `Ethnicity` for the ART for HIV dataset. *Equivalent classes* are then formed by combining these variables, resulting in groups such as `Male + Asian`

and `Female + African`. The risk associated with linking a synthetic patient s is estimated with

$$\frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{F_s}\times I_s\right), \tag{11}$$

where S represents the total number of records in the synthetic dataset, $I_s \in 0, 1$ equals one if the equivalent class of synthetic s is present in both datasets, and F_s denotes the cardinality of the equivalent class in the actual dataset.

The European Medicines Agency [45] and Health Canada [46] standards recommend that this risk should not exceed 9% to balance synthetic data utility and security. By following these measures, we ensure that our synthetic datasets are secure and suitable for public use.

E.4 Utility Investigation

E.4.1 RL Setup

We employ both the synthetic and real datasets to train RL agents, and we consider a synthetic dataset to achieve a high level of utility if an RL agent trained on both real and synthetic datasets generates similar actions when presented with clinical conditions of patients.

We partitioned each dataset into a set of observational variables and a set of action variables. The observational variables describe the clinical condition of a patient, while the action variables define the actions an RL agent could take. We adopt the approach in Liu *et al.* [74] to reduce the observational dimensionality to five variables using cross decomposition [75]. Next, we applied K-Means clustering [76] with 100 clusters to define the state space and assigned each data point to their corresponding cluster label. The action space was defined as the set of unique values of the action variables.

Subsequently, we employed published reward functions to determine the optimal actions that an RL agent should take given a patient state[3]. We select batch-constrained Q-learning [47] for utility investigation, and we update the policies for 100 iterations with a step size of 0.01.

E.4.2 CPH Setup

Utility studies assessing the effectiveness of synthetic datasets in healthcare settings are not only crucial but also under-represented in the literature. While these studies are often conducted within the context of modern machine learning paradigms such as RL, the importance of traditional statistical models should not be overlooked. In particular, CPH models continue to be pivotal tools in clinical research.

The HIV dataset captures the impact of various ART combinations on controlling viral load. The dataset contains key variables including ART types (*i.e.,* `FTC + TDF` vs `3TC + ABC`); and we manually deried the time required to achieve a `VL` of less than 1,000 copies/mL–a significant clinical milestone [23], and event status, which is categorised as either achieved or censored. CPH models for this study are constructed using the survival analysis package in R by Therneau & Grambsch [**?**].

To evaluate the utility of the ground truth $\mathfrak{D}_{\text{real}}$ against the GAN-generated $\mathfrak{D}_{\text{null}}$, and our DPM-simulated $\mathfrak{D}_{\text{alt}}$, we visualise the survival curves for each ART type. These visualisations facilitate a comparative assessment of the predictive accuracy across the datasets with regard to the time required to control viral load.

E.4.3 Setup for Inspecting Viral Suppression and Immune Recovery

In a quest to further evaluate the utility of synthetic datasets for HIV treatment via ART, we focus on quantifying patient recovery trajectories in both real and synthetic datasets. Adhering to WHO guidelines [23], we categorise patients into those initiating treatment with either the recommended first-line medications or alternative medications, as outlined in Table 4.3 on page 154/480.

[3]Refer to Gottesman *et al.* [59] and Parbhoo *et al.* [27] for the reward functions for acute hypotension and ART for HIV. In addition, see Sections 7.1 in the Appendix of Kuo *et al.* [5] for additional details on the implementation for acute hypotension; and likewise Section 4.3.5 in Kuo *et al.* [34] for ART for HIV.

In order to facilitate comparison, we examine these two distinct patient groups across a 60-month timeline, a period chosen for its alignment with the synthetic datasets $\mathfrak{D}_{\text{null}}$ and $\mathfrak{D}_{\text{alt}}$, which have fixed 60-month patient records, unlike the variable-length records in the real dataset $\mathfrak{D}_{\text{real}}$.

Monthly, we compute the percentages of patients in each group achieving the key benchmarks of viral suppression (VL $<$ 200 copies/mL) and immune recovery (CD4 $\geq$ 500 cells/μL). These metrics are then visualised to establish an initial comparative baseline.

F More Experimental Results

A series of hierarchical statistical tests was employed as shown in Figure 9 following the work of Kuo *et al.* [5]. Our objective was to determine whether the statistics of those data from the synthetic dataset used to train a neural network would be considered to be highly similar to the real dataset during iterative batch training. To achieve this, we sampled a batch of synthetic and real data with a batch size of 32 for a maximum of 100 iterations (hence the denominators in the Table for the tests are 100). We then performed all statistical tests along the variable dimension.

F.1 On ART for HIV

Variable Name	KS-Test	t-Test	F-Test	Three Sigma Rule Test
VL	45/100	94/100	45/100	100/100
CD4	94/100	92/100	77/100	99/100
Rel CD4	94/100	87/100	89/100	100/100
Gender	100/100	- -	100/100	- -
Ethnic	99/100	- -	100/100	- -
Base Drug Combo	97/100	- -	98/100	- -
Comp. INI	96/100	- -	100/100	- -
Comp. NNRTI	79/100	- -	97/100	- -
Extra PI	99/100	- -	100/100	- -
Extra pk-En	99/100	- -	100/100	- -
VL (M)	99/100	- -	100/100	- -
CD4 (M)	100/100	- -	91/100	- -
Drug (M)	99/100	- -	100/100	- -

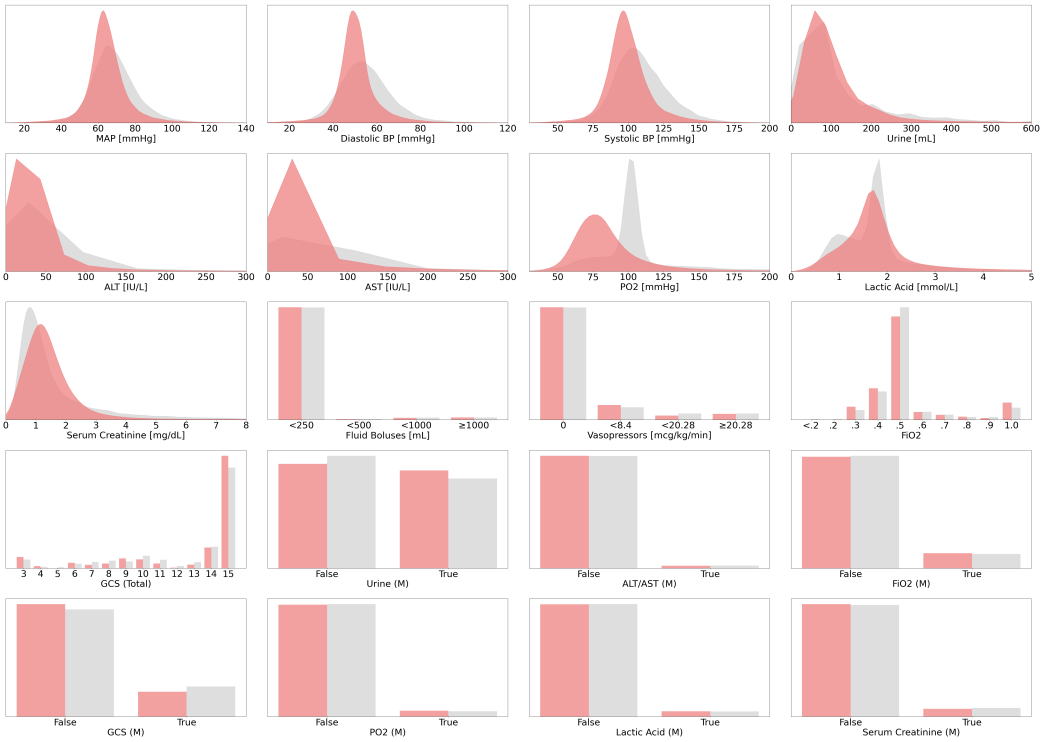
Table 4: Results on the hierarchical statistical tests for ART for HIV.

F.2 On Acute Hypotension

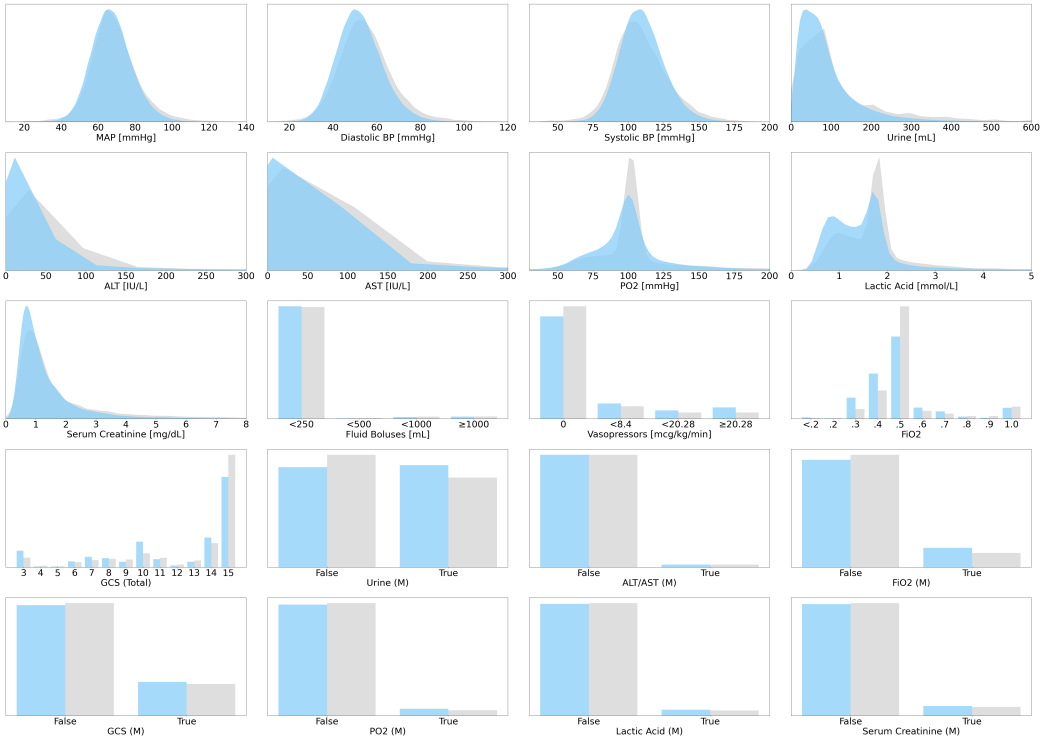
Dataset Realism
The KDE plots and barplots for the individual variable comparisons are presented in Figure 10. The grey bars represent the real variables from $\mathfrak{D}_{\text{real}}$, while the respective pink and blue bars in subplots 10(a) and 10(b) depict the synthetic variables in $\mathfrak{D}_{\text{null}}$ and $\mathfrak{D}_{\text{alt}}$, generated using Kuo *et al.* [5]'s Health Gym GAN and our DPM. Overall, the distributions in both subplots are comparable to their real counterparts in $\mathfrak{D}_{\text{real}}$. We observed that DPM captured the multi-modal nature of clinical variables better than GAN (*e.g.,* $\texttt{PaO}_2$ and Lactic Acid), but we also found that our DPM generated more instances of less common classes in $\texttt{FiO}_2$.

The synthetic variables in our DPM-generated hypotension dataset $\mathfrak{D}_{\text{alt}}$ are representative of their real counterparts in $\mathfrak{D}_{\text{real}}$. The statistics in Table 5 revealed that all variables passed the three sigma rule test and are reliable. Most variables passed the KS test and thus captured detailed information in the real distributions. The minority of variables that failed the KS test still passed the t-test and F-test, demonstrating that both the mean and the variance are captured and only missing the extreme details in the cumulative distribution function.

The correlations for acute hypotension are depicted in Figure 11. The panel on the left corresponds to the synthetic dataset $\mathfrak{D}_{\text{null}}$ generated by Kuo *et al.* [5]'s GAN; the middle panel represents our DPM-simulated dataset $\mathfrak{D}_{\text{alt}}$; and the panel on the right corresponds to the ground truth dataset $\mathfrak{D}_{\text{real}}$.



(a) Synthetic dataset $\mathfrak{D}_{\text{null}}$ from Kuo *et al.* [5] in pink.



(b) Synthetic dataset $\mathfrak{D}_{\text{alt}}$ from our DPM in blue.

Figure 10: Comparing the variables in acute hypotension, with those of $\mathfrak{D}_{\text{real}}$ in colour grey.

Variable Name	KS-Test	t-Test	F-Test	Three Sigma Rule Test
MAP	93/100	90/100	99/100	100/100
Diastolic BP	90/100	86/100	80/100	100/100
Systolic BP	93/100	95/100	91/100	100/100
Urine	88/100	87/100	98/100	99/100
ALT	54/100	91/100	87/100	97/100
AST	53/100	91/100	92/100	99/100
PaO2	46/100	89/100	76/100	99/100
Lactic Acid	29/100	85/100	97/100	98/100
Serum Creatinine	89/100	83/100	97/100	100/100
Fluid Boluses	100/100	- -	87/100	- -
Vasopressors	95/100	- -	89/100	- -
FiO_2	94/100	- -	95/100	- -
GCS	93/100	- -	86/100	- -
Urine (M)	98/100	- -	95/100	- -
ALT/AST (M)	100/100	- -	78/100	- -
FiO_2 (M)	94/100	- -	95/100	- -
GCS (M)	100/100	- -	98/100	- -
PaO_2 (M)	100/100	- -	94/100	- -
Lactic Acid (M)	100/100	- -	95/100	- -
Serum Creatinine (M)	100/100	- -	96/100	- -

Table 5: Results on the hierarchical statistical tests for acute hypotension.

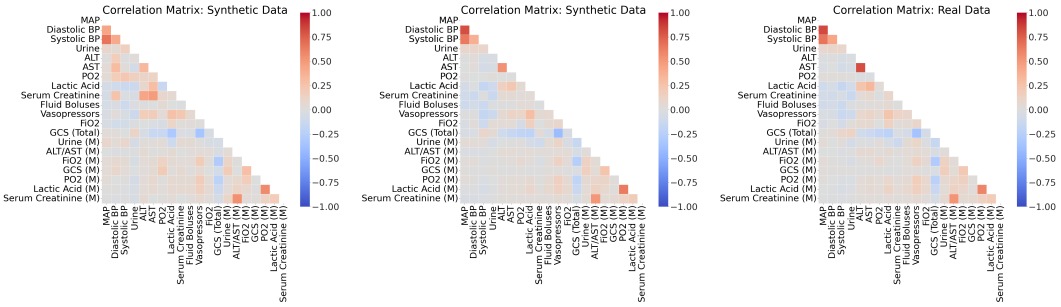


Figure 11: Comparing the correlations in acute hypotension.
The left panels depicts correlations in Kuo *et al.* [5]'s $\mathfrak{D}_{\text{null}}$. Whereas the middle and right panels respectively depict the correlations in our $\mathfrak{D}_{\text{alt}}$ and those in the ground truth $\mathfrak{D}_{\text{real}}$.

Figure 11 indicates that the correlations in our DPM-simulated dataset (located in the middle panels) exhibit a stronger resemblance to their real counterparts (located in the right panels) than those generated by GAN (located in the left panels).

	U (↓)	**CAT** (↑)
$\mathfrak{D}_{\text{null}}$ [5]	-2.1413	98.03%
$\mathfrak{D}_{\text{alt}}$ (ours)	**-2.4103**	**100.00**%

Table 6: Metric comparison.

Quantitative assessments on diversity and data structure are in Table 6. Category coverage (CAT) shows that all combinations of binary and categorical variables are present in our DPM-simulated dataset $\mathfrak{D}_{\text{alt}}$; but such is not the case for not all combinations in the GAN-generated dataset $\mathfrak{D}_{\text{null}}$ produced by Kuo *et al.* [5]. Moreover, the log-cluster metric (U) scores indicate that the latent structure embedded in our $\mathfrak{D}_{\text{alt}}$ is more realistic than that in $\mathfrak{D}_{\text{null}}$.

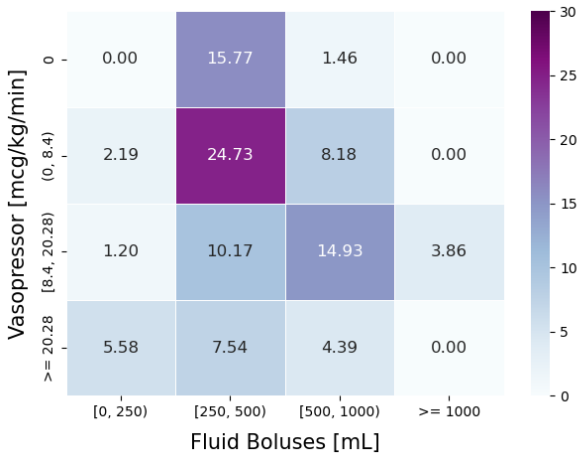
Risk Assessment

The variables in Table 3 are all related to the patient's bio-physiological states and do not contain any sensitive information that may reveal individuals' identities. Consequently, we only tested Euclidean distances and did not assess the disclosure risk. We found that records in our DPM-simulated synthetic dataset $\mathfrak{D}_{\text{alt}}$ matched none of those in the real hypotension dataset $\mathfrak{D}_{\text{real}}$. The minimum Euclidean distance between any synthetic record and any real record was 2.79 (> 0), indicating that no data was leaked into the synthetic dataset.

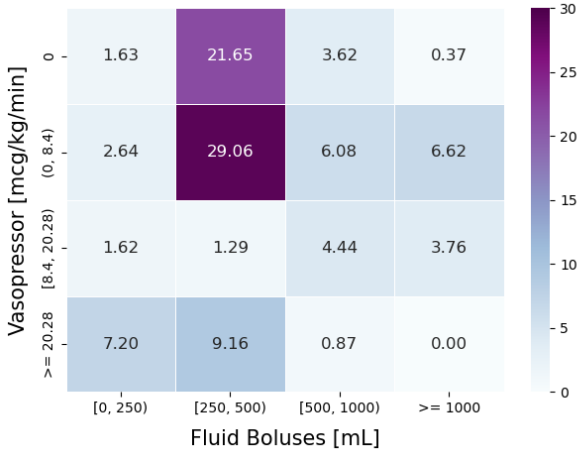
Utility

After training RL agents to suggest clinical treatments, we used heatmaps to visualise their action patterns. Each tile on the heatmap represents a unique action and its associated number indicates the frequency of that action as a proportion of all actions taken.

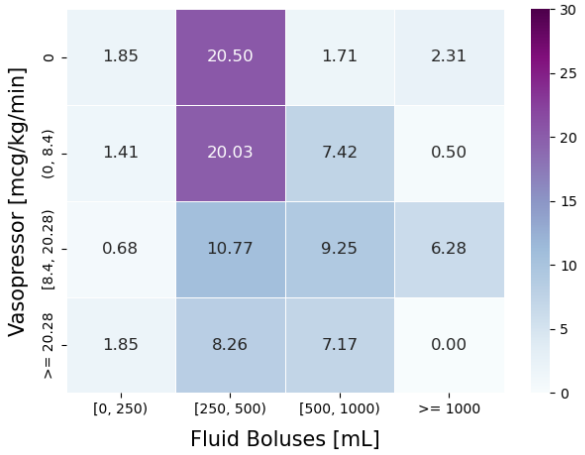
We depicted the action patterns of the RL agents for acute hypotension in Figure 12. The action space is spanned by `Vasopressor` and `Fluid Boluses`. Subplot (a) exhibits the actions taken by an RL agent trained on the real dataset $\mathfrak{D}_{\text{real}}$; subplots (b) and (c) respectively display the actions taken by RL agents trained on Kuo *et al.* [5]'s synthetic dataset $\mathfrak{D}_{\text{null}}$ and our DPM-simulated $\mathfrak{D}_{\text{alt}}$. The heatmap in subplot (c) shows a better alignment with its counterpart in subplot (a), indicating that the RL agent trained on our $\mathfrak{D}_{\text{alt}}$ suggested actions that were more similar to those suggested by the RL agent trained on $\mathfrak{D}_{\text{real}}$.



(a) RL policy trained on the real dataset $\mathfrak{D}_{\text{real}}$.



(b) RL policy trained on Kuo *et al.* [5]'s GAN-generated $\mathfrak{D}_{\text{null}}$.



(c) RL policy trained on our DPM-simulated $\mathfrak{D}_{\text{alt}}$.

Figure 12: Comparing the policies learned by RL agents on the acute hypotension datasets.
We illustrate the recommended policies of RL agents, trained using various acute hypotension datasets.
The RL action space is spanned by `Vasopressors` and `Fluid Boluses`.