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

Nicholas I-Hsien Kuo*, Louisa Jorm, Sebastiano Barbieri
Centre for Big Data Research in Health, University of New South Wales, Australia
*Corresponding email: n.kuo@unsw.edu.au

UNSW Centre for Big Data Research in Health

UNSW SYDNEY

Problem

Sensitive patient information are usually restricted, thereby limiting the access to data needed to develop effective machine learning models.

What is Already Known

Generative Adversarial Networks (GANs) can produce realistic synthetic surrogate clinical datasets, but suffer mode collapse, which significantly reduces the diversity and consequently undermines the utility of the data.

What this Study Adds

Diffusion Probabilistic Model (DPM) is a promising archetype of generative models that circumvents the practical training challenges in GANs. However, DPMs remain relatively under-explored in the research community.

Our Study

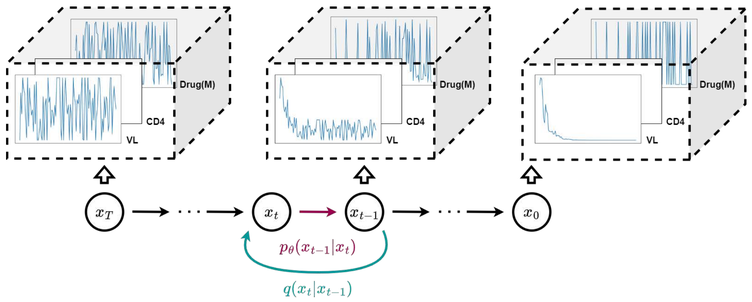
1) Extends the DPM application to simulate synthet longitudinal clinical data with mixed-type variables;
2) tested over 3 clinical conditions including acute hypotension, sepsis, and the ART for HIV;
3) validated the fidelity, the security, and the utility our synthetic datasets; and comparing the results with GAN-based SoTA.

Methods

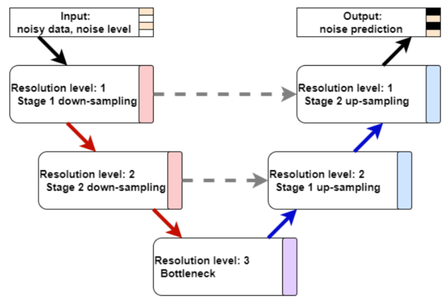
The DPM framework consists a forward diffusion process to remove distinguishable features and a reverse diffusion process to learn to recover data as if they were sampled from the real database.

We employ U-Net as our backbone model to extract meaningful information from noisy data while preserving its underlying structure.

The U-Net consists multiple layers of one-dimensional convolutional neural networks, allowing the extraction of high-level features at various levels of resolution.



The concept of the DPM framework



The U-Net backbone for denoising

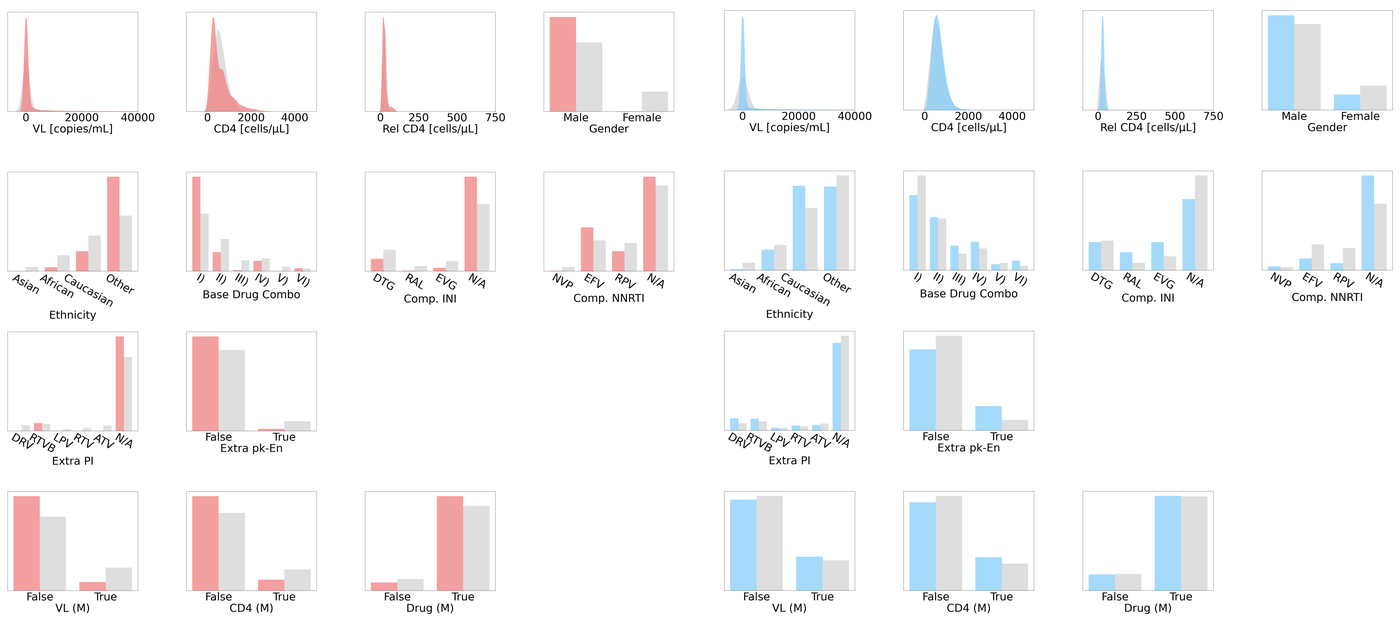
Critical Findings

1) Overall, DPM-simulated datasets are more realistic than GANs
2) DPMs are easier to train than GANs
3) DPMs do not suffer from mode collapse
4) DPMs generate categorical/binary variables with better representations
5) GANs generate numeric variables with lower bias (in both mean and variance)
6) GANs generate/sample synthetic data more efficiently than DPMs

Future work

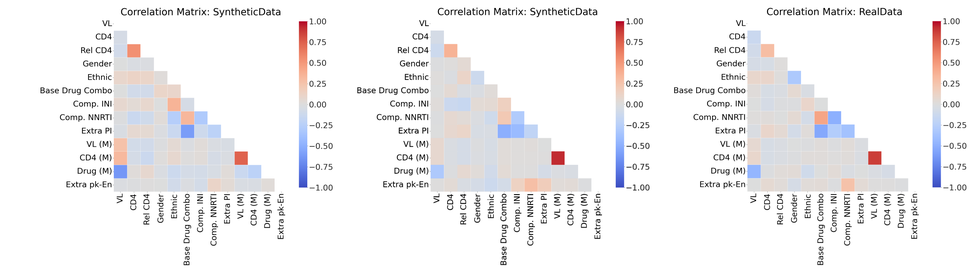
1) Design a DPM that generates less bias in numeric variables
2) Conduct a large scale utility study on the DPM-simulated dataset, to verify that the synthetic dataset is capable of substituting the ground truth for developing logistic regression, random forest, and deep learning algorithms such as reinforcement learning

Selected Results

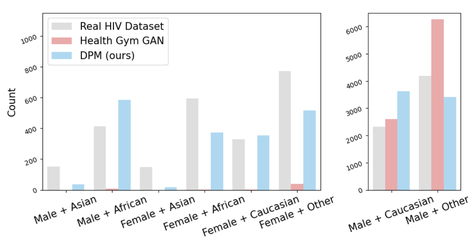


Variable distributions generated using a baseline WGAN-GP

Variable distributions generated using our novel DPM



Correlations from WGAN-GP from our DPM from the Real Data



Combinations of patient demographics

See more results in our paper regarding patient exposure risk and synthetic dataset utility.

Preprint: arXiv:2303.12281

Github: https://github.com/Nic5472K/ScientificData2021_HealthGym

Follow us on: HealthGym.ai

HealthGym.ai