2nd PANHELLENIC CONGRESS OF MEDICAL PHYSICS
4-6 OCTOBER 2024| EUGENIDES FOUNDATION

Creating virtual patients with a generative artificial intelligence algorithm for clinical studies

Anastasios Nikolopoulos[1,2], Vangelis D. Karalis[1,2]

[1]Department of Pharmacy, National and Kapodistrian University of Athens, Athens, Greece

[2]Institute of Applied and Computational Mathematics, Foundation for Research and Technology Hellas (FORTH), Heraklion, Greece

1. Background-Aim

Clinical trials often face challenges with limited sample sizes, which can affect the accuracy of results and raise ethical concerns due to human involvement *(Figure 1)*. A well-calculated sample size is essential, as a smaller sample can lead to unreliable outcomes, while a larger sample increases costs and patient exposure.

To address these issues, artificial intelligence (AI) offers innovative solutions, particularly in data augmentation. Our study explores the use of **Wasserstein Generative Adversarial Networks (WGANs)** to generate virtual subjects from small datasets, reducing the need for large samples in clinical trials.

By generating synthetic data, this approach aims to *minimize patient exposure, lower costs, and improve the efficiency of clinical research*.

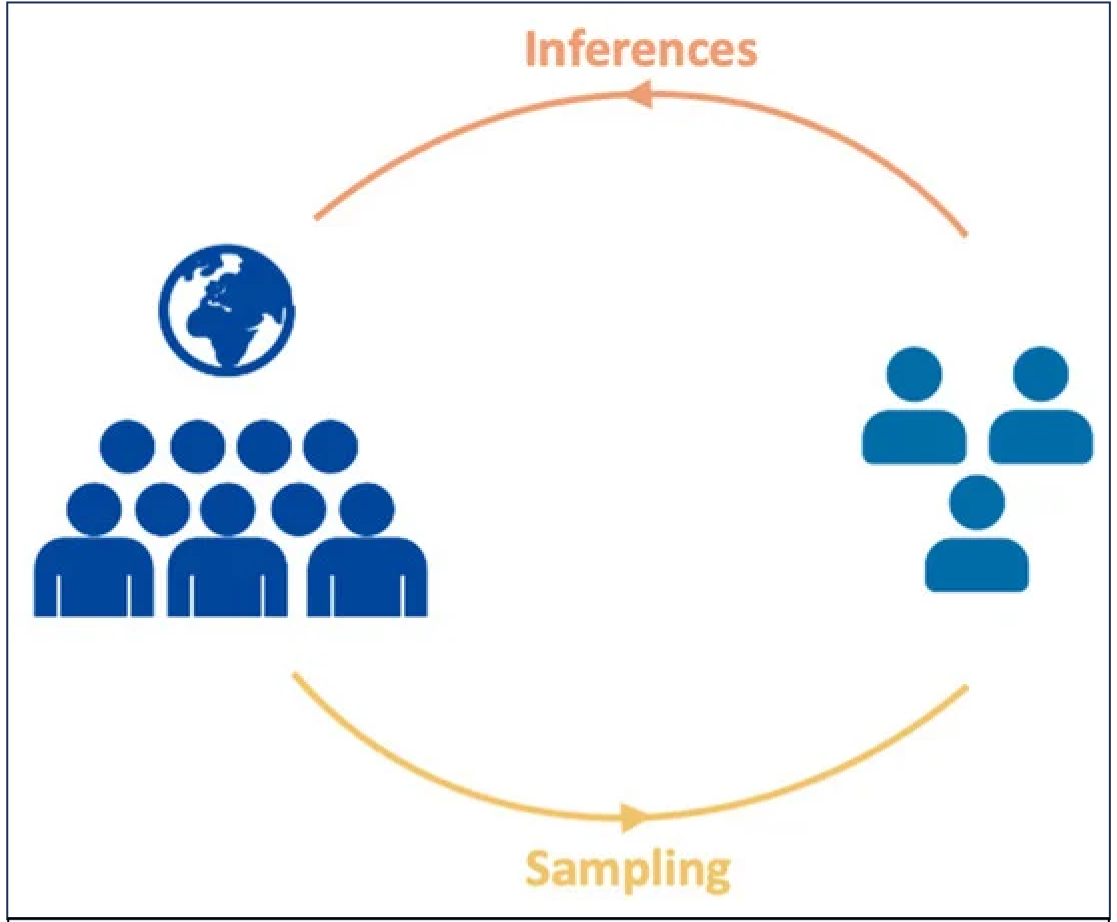

***Figure 1**: Illustration of the current clinical trial workflow.*

2. Materials & Methods

In clinical trials, an accurate sample size is crucial to assess treatment efficacy and safety. Traditional methods require large samples to ensure statistical power and reduce false negatives errors. The *steps of our research* included *(Figure 2)*:

1. Original Dataset: Create a population of **10,000 patients** (with Monte Carlo simulations).
2. Sample Dataset: *Randomly* select a small subset (**0.5% or 1% of the population**).
3. Generated Dataset: Use WGANs to create **virtual patients** based on the sample.
4. Comparison: **Assess the generated dataset's performance** by comparing it with the original and sample datasets.

The *goal is to demonstrate that the WGAN-generated data closely mirrors the entire population*, enhancing statistical power.

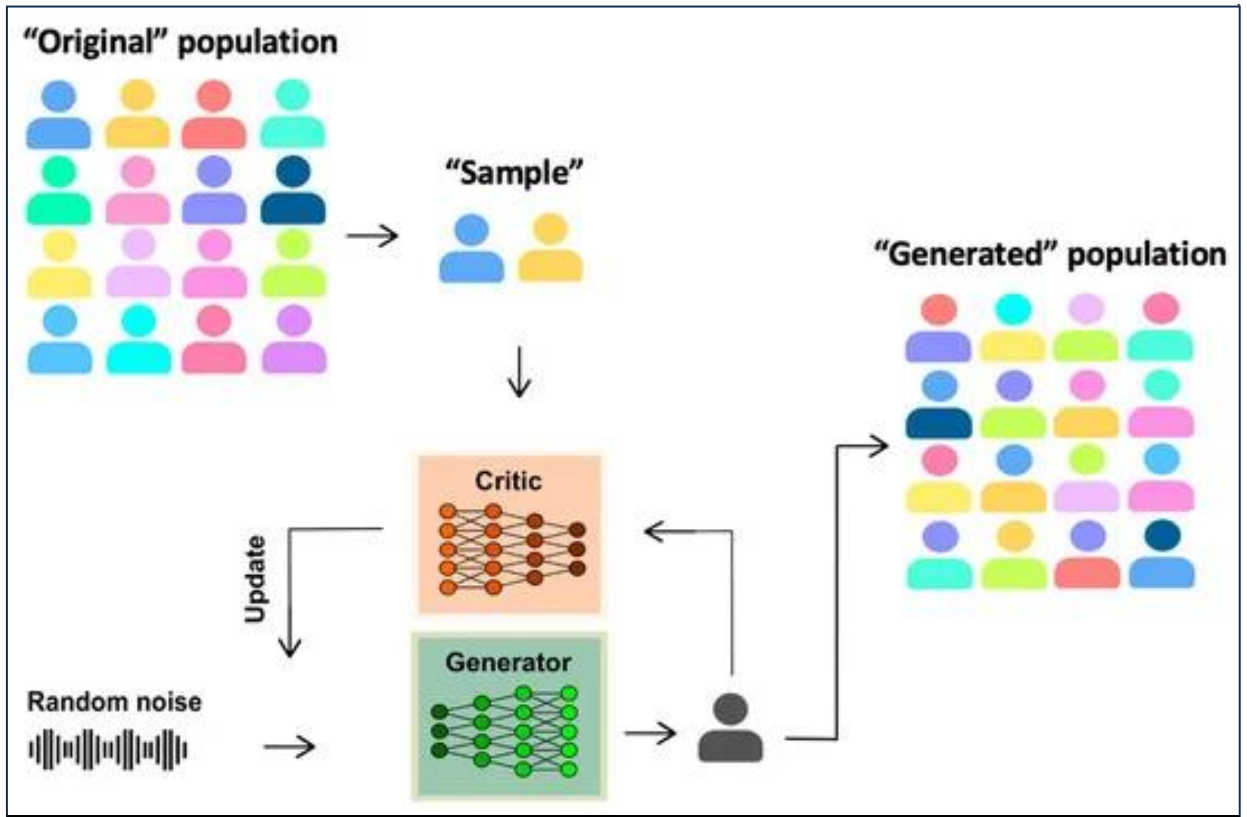

Figure 2*: Illustration of the proposed method for applying WGANs in clinical studies.*

2. Materials & Methods

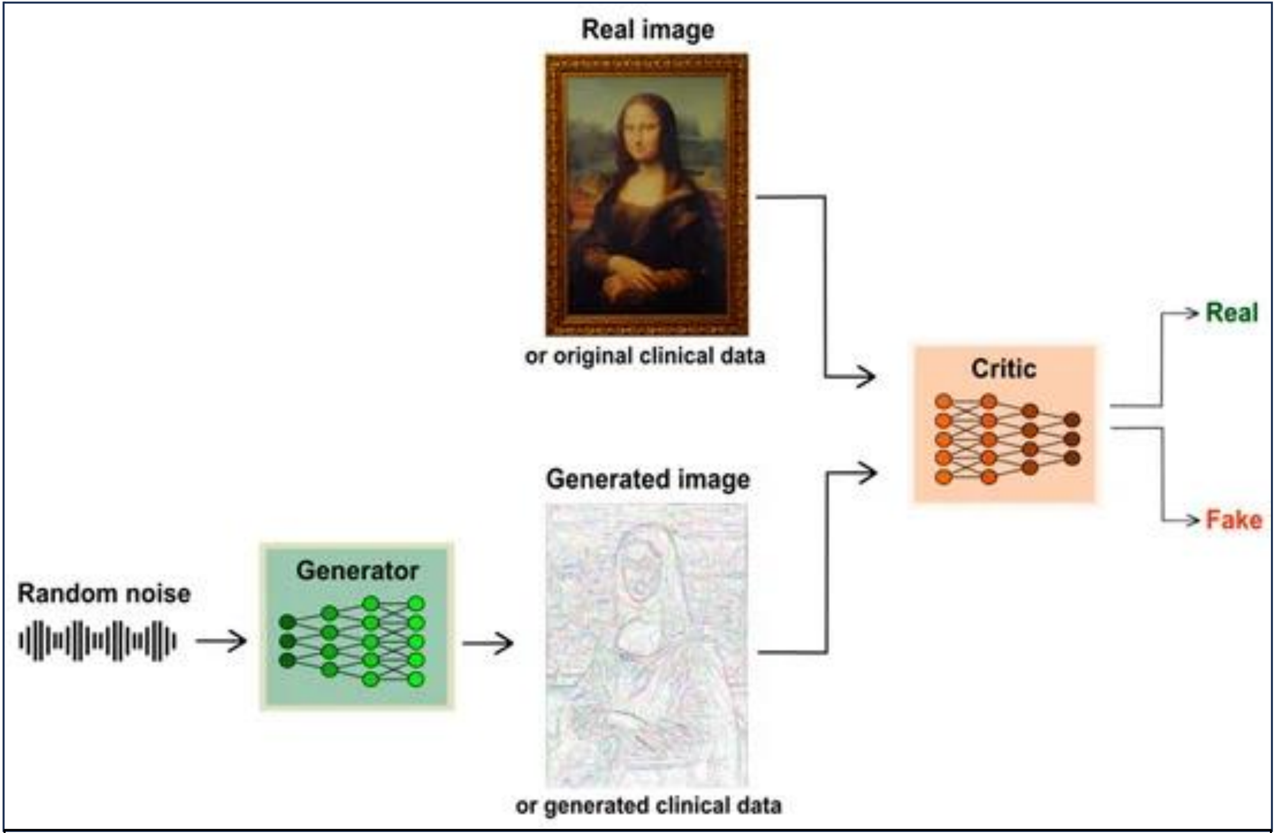

***Figure 3**: General depiction of WGAN operation. The diagram can be applied across various contexts, including clinical data, as all types of input, whether clinical or image-based, are ultimately converted into numerical data.*

WGANs, are used to generate realistic synthetic data, consisting of *two main parts*:

- **Generator**: Creates new data based on random noise.
- **Discriminator (Critic)**: Evaluates real vs. generated data.

Through training, the generator learns to produce *data that resembles the real dataset*, while the critic tries to *distinguish between real and fake data*. This back-and-forth competition improves the generated dataset over time, making it statistically similar to the original population *(Figure 3)*.

WGANs use a **Wasserstein distance metric** to ensure stable and meaningful convergence, addressing traditional GANs' issues with instability. In this study, WGANs generate **virtual clinical trial participants**, allowing *small sample sizes to yield reliable results*.

3. Results

In our study, we initially worked with a *0.5% sample rate (50 patients)* from a population of 10,000. Despite the small sample size, the generated dataset closely mirrored the original population’s distribution *(Figure 4)*, demonstrating **WGAN's ability to overcome sample size limitations**. This was consistent even with a higher Test/Reference (T/R) ratio. Additionally, when comparing the similarity between datasets, the **WGAN-generated data consistently showed higher resemblance to the original population than the sample dataset** *(Figure 5)*.

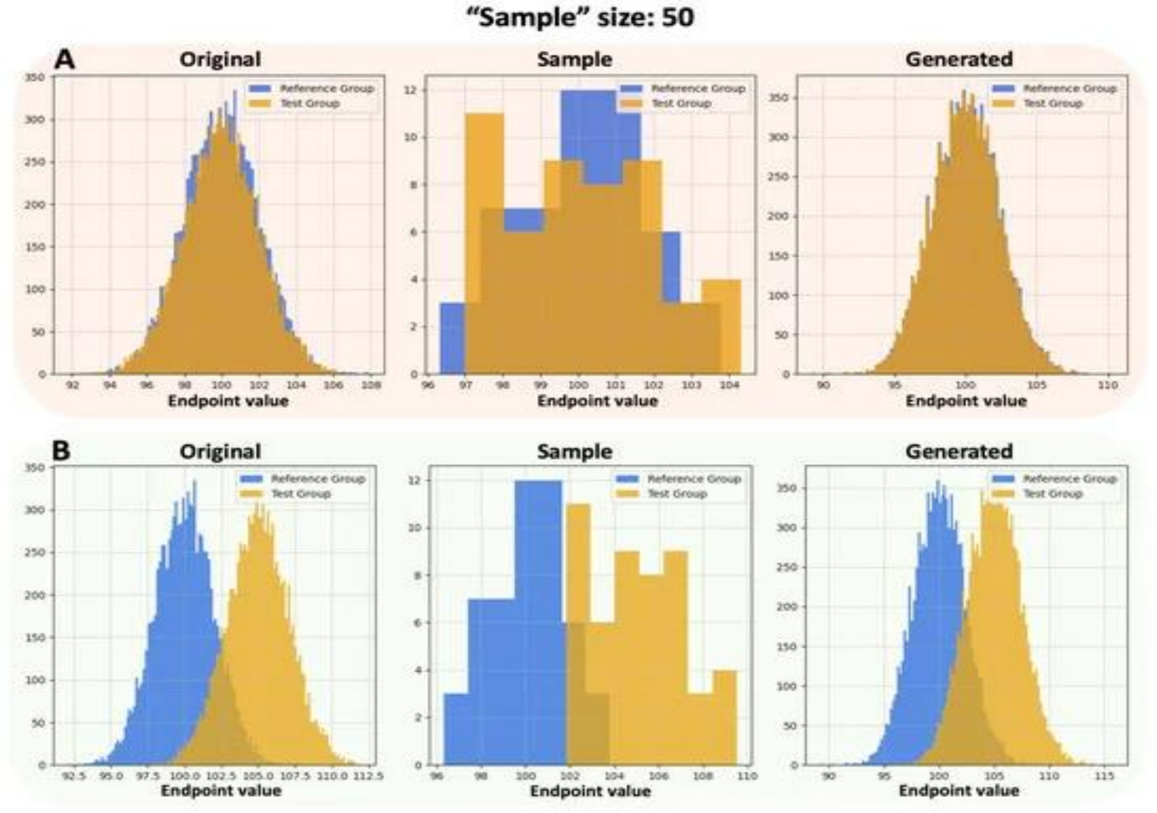

***Figure 4**: Distribution of the "original" population, the "sample" of 50 subjects, and the WGAN-generated dataset. The between-subject variability has a 20% coefficient of variation, with the T/R ratio set to 1.00 in panel A and 1.05 in panel B.*

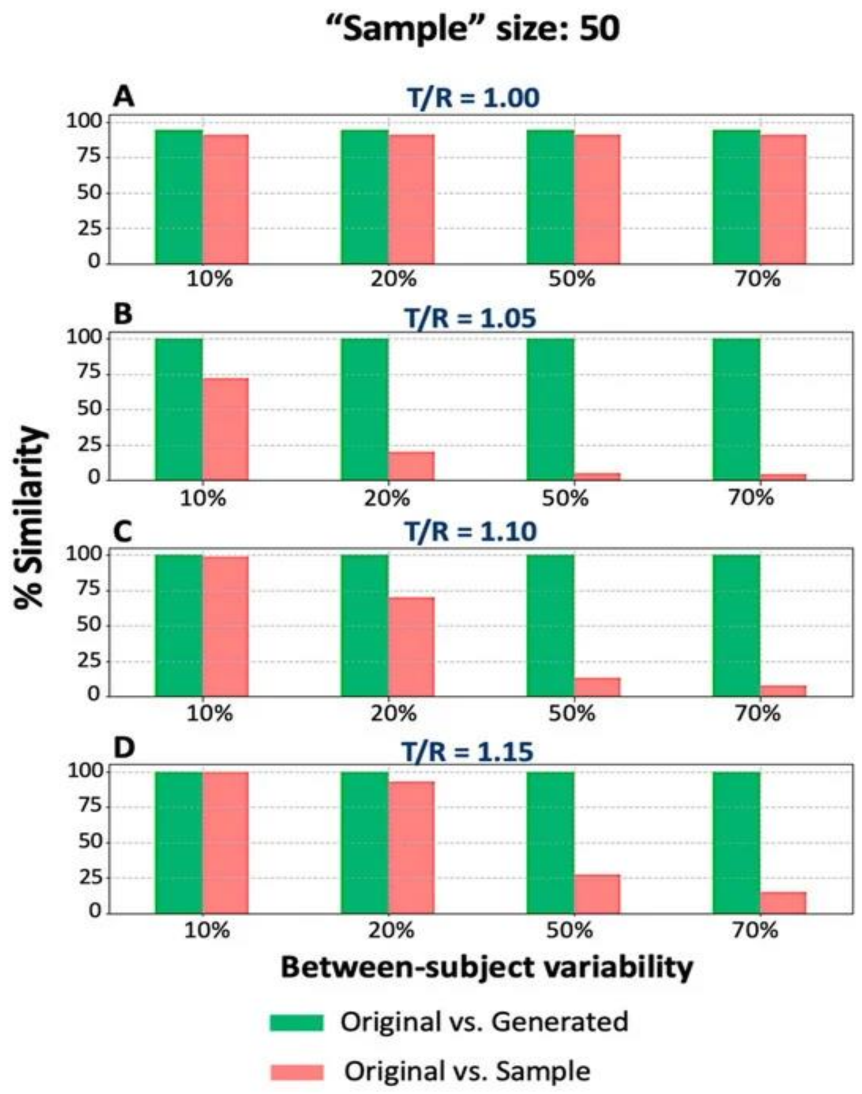

***Figure 5**: Percent similarity in performance between the "original," "sample," and "generated" datasets. The average T/R ratios are 1.00 (A), 1.05 (B), 1.10 (C), and 1.15 (D).*

3. Results

Further analysis was performed with a 1.0% sample rate (100 patients), yielding similar results, where **WGANs outperformed the sample in mimicking the original population**. Across different T/R ratios and variability levels, the generated data exhibited stable performance, with a high degree of similarity to the original population, regardless of the sample or generated size *(Figures 6-7)*.

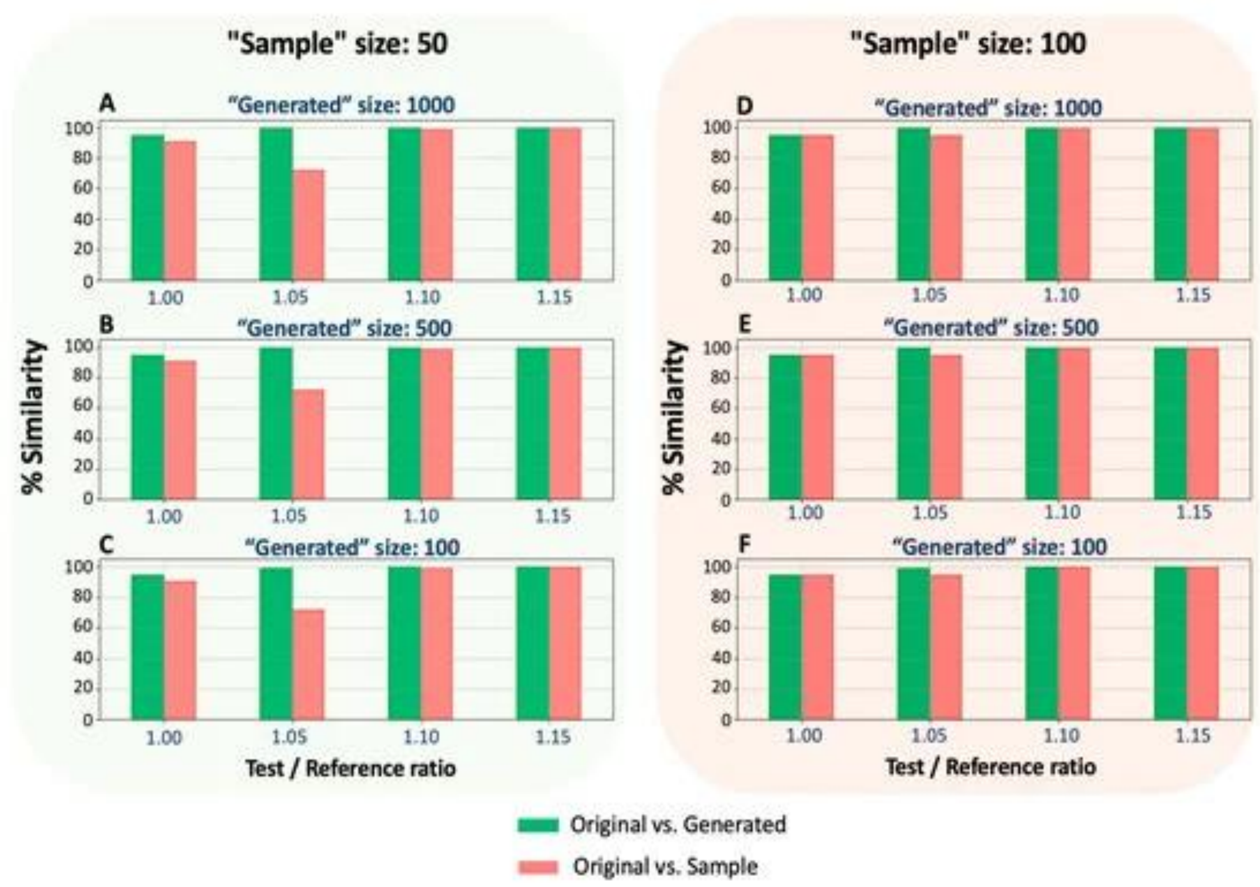

***Figure 6**: Percent similarity in performance among the "original," "sample," and "generated" datasets, with various generated and sample sizes. All scenarios involve a 10% between-subject variability.*

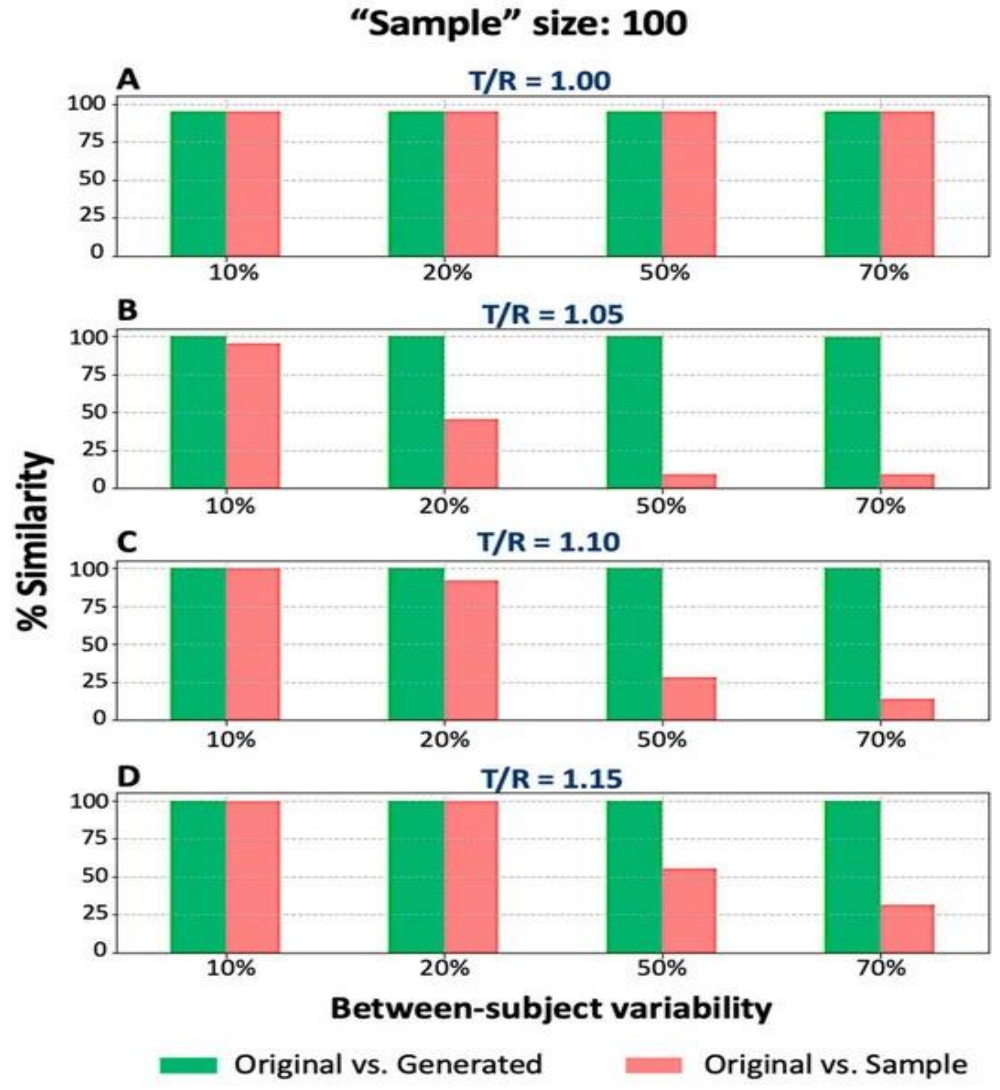

***Figure 7**: Percent similarity in performance among the "original," "sample," and "generated" datasets with the "sample" size set at 100, for various T/R ratios and between-subject variability values.*

4. Conclusions

This study introduces a **novel method for data augmentation in clinical trials using WGANs** to generate virtual subjects. By training WGANs on a small sample, we can create a synthetic population that mirrors the statistical properties of the entire population, enhancing the study's power without increasing participant numbers.

Our research demonstrated that the **WGAN-generated data consistently performed better** than traditional small samples across various scenarios. This approach offers a promising solution to reduce clinical trial costs, time, and human involvement, while ensuring accuracy and reproducibility in results. However, real-world clinical data and ethical guidelines are necessary to further validate this method.

You can find more about our paper, here: