

Article

Implementation of a Generative AI Algorithm for Virtually Increasing the Sample Size of Clinical Studies

Anastasios Nikolopoulos^{1,2}  and Vangelis D. Karalis^{1,2,*} 

¹ Department of Pharmacy, School of Health Sciences, National and Kapodistrian University of Athens, 15784 Athens, Greece; anastnik@pharm.uoa.gr

² Institute of Applied and Computational Mathematics, Foundation for Research and Technology Hellas (FORTH), 70013 Heraklion, Greece

* Correspondence: vkaralis@pharm.uoa.gr; Tel.: +30-210-727-4267

Featured Application: This work proposes the application of AI generative algorithms, specifically Wasserstein Generative Adversarial Networks (WGANs), to reduce the sample size in clinical trials. Additionally, a novel methodological procedure is established for this study, where the entire population, a sample, and AI-synthesized data are compared through Monte Carlo simulations. It is suggested that utilizing only a small subset of the true population along with WGANs can yield results similar to those obtained from the entire population.

Abstract: Determining the appropriate sample size is crucial in clinical studies due to the potential limitations of small sample sizes in detecting true effects. This work introduces the use of Wasserstein Generative Adversarial Networks (WGANs) to create virtual subjects and reduce the need for recruiting actual human volunteers. The proposed idea suggests that only a small subset (“sample”) of the true population can be used along with WGANs to create a virtual population (“generated” dataset). To demonstrate the suitability of the WGAN-based approach, a new methodological procedure was also required to be established and applied. Monte Carlo simulations of clinical studies were performed to compare the performance of the WGAN-synthesized virtual subjects (i.e., the “generated” dataset) against both the entire population (the so-called “original” dataset) and a subset of it, the “sample”. After training and tuning the WGAN, various scenarios were explored, and the comparative performance of the three datasets was evaluated, as well as the similarity in the results against the population data. Across all scenarios tested, integrating WGANs and their corresponding generated populations consistently exhibited superior performance compared with those from samples alone. The generated datasets also exhibited quite similar performance compared with the “original” (i.e., population) data. By introducing virtual patients, WGANs effectively augment sample size, reducing the risk of type II errors. The proposed WGAN approach has the potential to decrease costs, time, and ethical concerns associated with human participation in clinical trials.

Keywords: sample size; clinical trials; generative AI algorithms; WGANs; data augmentation



Citation: Nikolopoulos, A.; Karalis, V.D. Implementation of a Generative AI Algorithm for Virtually Increasing the Sample Size of Clinical Studies. *Appl. Sci.* **2024**, *14*, 4570. <https://doi.org/10.3390/app14114570>

Academic Editor: Qi-Huang Zheng

Received: 23 April 2024

Revised: 21 May 2024

Accepted: 24 May 2024

Published: 26 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Determining the appropriate sample size is of pivotal concern in clinical studies, since small sample sizes may be incapable of detecting true effects [1]. Not only does sample size impact accuracy and reliability but it also directly influences the associated costs and time investments in a clinical trial. It should also be noted that the participation of humans in the studies raises ethical concerns. Additionally, the size of the sample is a crucial determinant in deciding whether to approve or reject the results of clinical trials, regardless of the clinical efficacy or inefficacy of the tested drug [2].

Given the logistical challenges of studying entire populations, clinical studies often rely on sampling. Findings derived from samples aim to be applicable to the broader

population, sometimes extending to future contexts. Therefore, it is crucial for the sample to accurately represent the population, a task facilitated by using appropriate sampling methodologies [3]. The determination of sample size aligns with the hypothesis and study design. The fundamental factors for estimating sample size in clinical trials include type I error (i.e., alpha), desired statistical power, measured endpoint(s), variability of the endpoint(s), limits of acceptance, etc. [4]. The sample size must be appropriate—neither excessively large nor too small. An excessively large sample would unnecessarily require the enrollment of more patients than required for the study objectives, which is unethical. Conversely, a sample size smaller than necessary would lack sufficient statistical power to address the primary research question, potentially leading to statistically nonsignificant results solely due to an inadequate sample size (type II or false negative error) [3,5–10].

To overcome the constraints imposed by limited sample sizes and reduce human involvement, computational tools can be utilized to expand the sample size using artificial methods [10]. Another strong tool that has emerged in recent years in this direction is the use of artificial intelligence. The number of AI applications has been increasing exponentially in recent years. In healthcare, there are numerous use cases of AI and machine learning, including drug discovery, medicine, dentistry, anesthesiology, and ophthalmology [11–16]. One recent application of AI proposed by our research group is data augmentation, which involves virtually increasing a sample by generating new data from existing data [17,18]. In this context, several other studies have evaluated the effectiveness of diverse augmentation approaches where the Wasserstein Generative Adversarial Networks (WGANs) have exhibited superior performance compared with other methods [19].

Introduced by Goodfellow [20], Generative Adversarial Networks (GANs) are a type of generative model capable of producing novel content based on some training data. GANs find application in various medical fields, including oncology for developing new molecules and enhancing image resolution [21–23]. However, their most prevalent application is in generating new images. Comprising two Artificial Neural Networks (ANNs), the generator and the discriminator, GANs operate by pitting these networks against each other: the generator generates new data instances, while the discriminator assesses their authenticity [24]. WGANs enhance traditional GANs by utilizing the Wasserstein distance, also recognized as the earth mover's distance, as the objective function during training. Wasserstein distance improves model stability, while mitigating issues like mode collapse and convergence. This distance metric offers a more insightful evaluation of the resemblance between real and generated data distributions compared with the probability-based metrics utilized in conventional GANs. In WGANs, the discriminator is called a "critic" because its role goes beyond binary classification but evaluates the quality of generated samples relative to real ones [25,26].

In this study, we present a novel approach aiming to reduce large sample size needs in clinical trials. The proposed approach suggests training an artificial neural network, specifically a WGAN, on a limited dataset (the "sample") and then utilizing a WGAN to generate virtual subjects. The proposed approach suggests that only the small subset (i.e., the "sample") of the true population can be used along with WGANs to create a virtual population (the "generated" dataset). To demonstrate the applicability of the proposed approach, Monte Carlo simulations of clinical studies were performed to compare the performance of the WGAN-synthesized virtual subjects (i.e., "generated" dataset) against both the entire population (the so-called "original" dataset) and a subset of it, namely, the "sample". The ultimate objective of this work is to illustrate the following:

- How AI techniques can be applied to generate virtual volunteers/patients, as an effort to reduce the need for large numbers of real patient data.
- The introduction of "virtual patients" to decrease the costs and times of clinical studies as well as reduce the human exposure in clinical studies.

Overall, it aims to offer a potential solution to the challenges associated with small sample sizes and human exposure in clinical research.

2. Materials and Methods

Artificial intelligence involves steps such as data acquisition, designing effective systems for data utilization, deriving precise or approximate conclusions, and implementing self-corrections or adjustments [27]. Machine learning and deep learning are two subfields of AI where the first one creates algorithms for data analysis to identify characteristics within datasets, while the latter one represents a contemporary advancement of traditional neural network methods. Deep learning can be conceptualized as a neural network comprising numerous layers. The advancement in modern computing has facilitated the construction of deep neural networks with a high volume of layers, a feat unattainable with conventional neural networks [27]. Certain machine learning models have demonstrated capabilities to achieve results comparable to, and in certain cases, even superior to human performance levels [28]. In pharmaceutical research, AI finds application across various stages of drug discovery and development, including formulation development for drug delivery, drug modeling, dosage design, protein structure analysis, and pharmacokinetic and pharmacodynamic modeling, as well as clinical trials [27].

2.1. Strategy of the Analysis

In clinical trials, determining an appropriate sample size is crucial for accurately assessing the effectiveness and safety of treatments (e.g., a new medicine against the existing reference treatment). Hypotheses related to the study drug are formulated based on primary endpoints and are evaluated using statistical tests within an explicitly defined clinical design to ensure accurate and reliable results (Figure 1). It is essential to clearly state these assumptions when calculating sample size, as each hypothesis requires specific sample size considerations to achieve desired statistical assurance (e.g., at most 5% false positive and 20% false negative) [4,29].

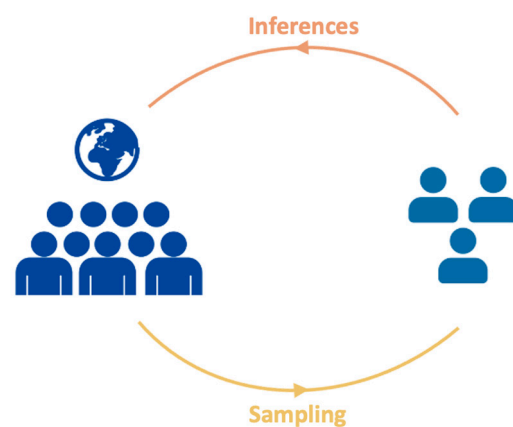


Figure 1. Visual representation of the existing clinical trial process.

To demonstrate the suitability of the WGAN-based approach, a new methodological procedure needs to be established and applied. The purpose of this work is to present a novel concept for reducing the sample size needed for clinical trials by using an appropriate ANN like WGANs. Specifically, the research idea involves the following steps (Figure 2):

1. Create an entire population (“the original dataset”).
2. Take a subset (the “sample”) of the population and conduct a clinical study using only this limited number of participants.
3. Combine the “sample” information and the WGANs to create virtual subjects, namely, the “generated” dataset, thus increasing statistical power.

The aim is to achieve high statistical power while avoiding an increase in the false positive rate. In essence, the purpose is to draw conclusions from a small sample in a clinical trial that mirror those that would be drawn from the entire population. Since achieving this through classical statistical methods is unfeasible, a generative AI algorithm, specifically

WGANs, is used to address this issue. To demonstrate the applicability of the new approach, it is important to show that the performance of the combination small sample/WGANs that leads to the “generated” dataset closely resembles that of the entire population (i.e., the “original” dataset), and notably exceeds the performance of the “sample” dataset (Figure 2). It should be underlined that performance pertains to the statistical power of the study and type I error, not safety considerations.

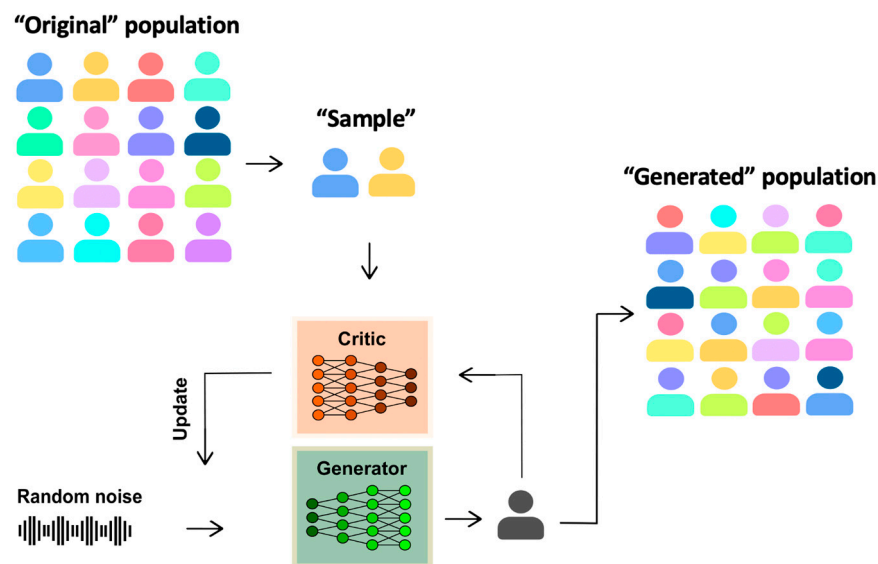


Figure 2. Proposed approach for the implementation of WGANs in clinical studies.

This study compares the performance of the WGAN-“generated” dataset against that of the entire population (i.e., the “original”) and the “sample” datasets. Ideally, the “generated” dataset performance should align closely with that of the “original”, indicating that the desired statistical properties can be achieved using only a small sample of actual subjects. If this approach is shown to hold, then it offers numerous benefits including cost-effectiveness, earlier trial completion, and simplification of complexities.

To demonstrate the usefulness of WGANs, an *in silico* experimental approach was implemented in this study. Below is a brief overview of the procedure, which was performed in a Jupyter Lab environment with Python:

- i. Monte Carlo simulations were used to virtually create 10,000 patients, which refer to the entire population, namely the “original” dataset. To simulate the conditions of two interventions (e.g., two different medicines), two distributions were created with 10,000 patients each.
- ii. Each distribution had a mean of the hypothetical endpoint equal to 100 units. Also, several between-subject variability values were used for the random generation of the populations: 10%, 20%, 50%, and 70%.
- iii. Two distributions were created, referring to the Test (T) and Reference (R) treatment. In addition, several T/R ratios among the means of the distributions were examined.
- iv. The T and R values were preprocessed via standard scaling.
- v. To obtain the “sample”, we mimic the process used in real-world scenarios, namely, random sampling from the “original” dataset. We utilized various proportions of the “original” dataset, ranging as low as 0.5% and 1.0%, corresponding to 50 and 100 subjects in the “sample”, respectively.
- vi. The WGANs were applied in the “sample”, which led to the synthesized values, which form the “generated” dataset (Figure 2). The size of the synthesized data initially was set equal to the population size, namely, 10,000 subjects. Later in this analysis, smaller sizes of the “generated” dataset were used, such as 1000, 500, and 100.

- vii. To simplify the process and avoid complexities, a simple statistical criterion (t -test) was applied to compare the T and R populations. Pairwise comparisons were then made among the “original”, “sample”, and “generated” datasets to assess their performance.
- viii. Each success (or failure) of the t -test was recorded for all datasets.
- ix. Since the “original” data represents the entire population, reflecting the true results, an additional comparison metric was introduced to assess the analysis outcomes. This metric, termed “similarity”, gauges the concordance of statistical results between the “sample” and “generated” datasets in comparison to the “original” dataset. During each repetition of the trials, we recorded whether the statistical outcomes derived from analyzing the “sample” and “generated” datasets agreed with those obtained from the “original” population. This comparison aimed to evaluate the fidelity of the generated datasets in capturing the underlying characteristics of the original population. The concept of “similarity” encompasses the simultaneous success or failure of the statistical evaluation, translating into the degree of resemblance between the distributions of the “sample” and “generated” datasets with the “original” population. A high “similarity” rate indicates that the statistical properties of the generated datasets closely mirror those of the original population, which is desirable for the validity and reliability of the generated data.
- x. The above steps (i–vii) were repeated many times (5000) to obtain robust estimates. Statistical power for each case was calculated as the number of successes after all repetitions divided by the total number of repetitions.

The ideal architecture of the WGANs and the fine-tuning of hyperparameters were exhaustively examined following the preceding step ‘vi’. Additionally, seeds were set in all stochastic processes to ensure reproducible results across different program executions. Once the final hyperparameters were determined, a variety of scenarios were executed, which remained consistent throughout the entire experimental procedure. These measures were implemented to ensure the reproducibility of the entire framework.

2.2. Wasserstein Generative Adversarial Networks

A GAN is an advanced machine learning tool for creating realistic synthetic objects. It comprises two neural networks, the generator and the discriminator, engaged in a competitive game [30]. The generator generates data resembling originals, while the discriminator identifies fakes. Through training, both networks improve until the generated objects are indistinguishable from originals [30]. WGANs address stability and convergence issues in training by using weight clipping to enforce a 1-Lipschitz constraint. This constrains network weights, promoting smaller, more stable updates and enhancing convergence. Compared with traditional GANs, WGANs use the Wasserstein distance for a more meaningful measure of similarity between real and generated data distributions [25].

It is critical to recognize that WGANs can capture complex data distributions (shapes) and generate samples that closely resemble real data, whereas, for example, a normal distribution generates random samples based on a specified mean and standard deviation without necessarily capturing the underlying structure of the data. Essentially, WGANs aim to learn the underlying data distribution and generate samples that adhere to it, whereas a specific distribution (e.g., normal) can only provide random samples based on the predefined parameters (mean and variance). Furthermore, the flexibility of GAN architectures plays a critical role in their application. Both the generator and discriminator, within a GAN, can be customized to accommodate specific shapes and dimensions of input and output data. This adaptability stands in stark contrast to bootstrapping, which, as a resampling technique, is used primarily to estimate the distribution of a statistic over a sample and maintains the exact same underlying distribution and characteristics as the original dataset. This inherent adaptability of GANs makes them suitable for a wide range of data types and formats. In our methodology, we exploit this flexibility to ensure that our GAN model is optimally designed to handle the specific characteristics of the data we aim to generate and analyze.

Figure 3 illustrates the operation of WGANs; initially, the generator takes random noise as input and produces an object, such as a value or an image. Subsequently, the critic evaluates both the real object and the generated one to discern between real and fake. The feedback from the critic is then used to update the generator, iteratively improving its ability to generate realistic objects. The goal is for the generator to create objects that are indistinguishable from real ones, ultimately achieving a state where the critic cannot differentiate between real and generated objects. The image labeled as a “real image” in Figure 3 can be substituted with original data. Using the “generating image” (generated data), the discriminator (critic) can then determine whether to classify it as real or fake.

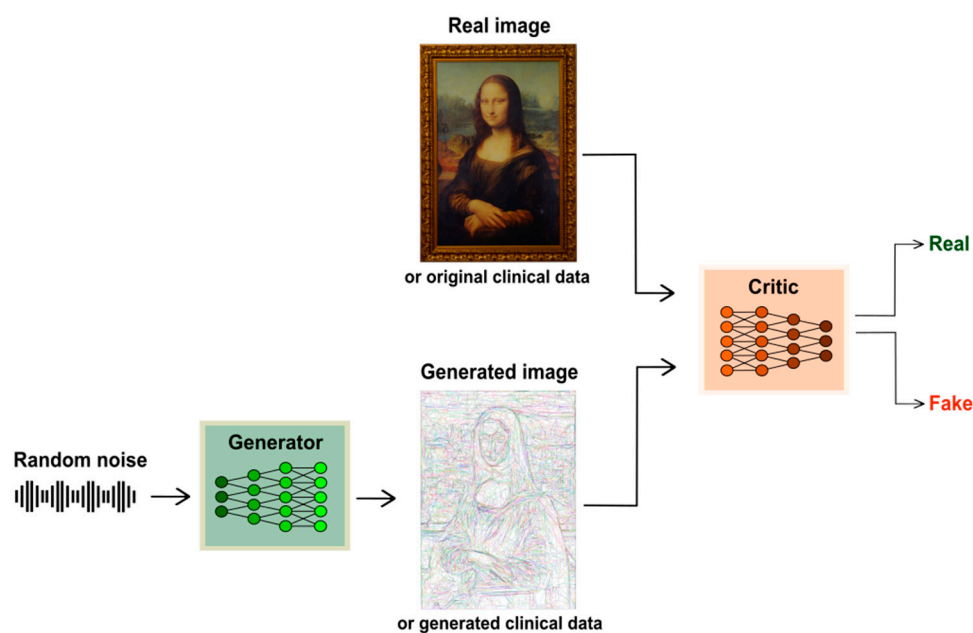


Figure 3. General representation of the operation of WGANs. This figure adopts a visual approach designed to facilitate an intuitive understanding of the WGAN algorithm workflow. By transcending traditional graphical representations, we aim to make the concepts accessible to readers with diverse backgrounds and varying levels of expertise. The diagram is universally applicable, as visual data representations can extend to clinical data scenarios and implementing original data, given that all types of input (such as clinical and image data) are ultimately translated into numbers.

In the adversarial modeling framework, both generator and discriminator models use multilayer perceptrons. The generator G maps noise z to data space, while the discriminator D distinguishes between real data and generated data $G(z)$. During training, D learns to maximize the probability of correctly classifying samples, while G learns to minimize the probability of being classified incorrectly. This process forms a two-player game where D aims to maximize and G aims to minimize the function (Equation (1)) [20]:

$$V(D, G) = E_{x \sim p_{data}}(x)[\log D(x)] + E_{z \sim p_z}(z)[\log(1 - D(G(z)))] \quad (1)$$

In simpler terms, D tries to tell real from fake, and G tries to generate data that fools D , resulting in a balance between the two networks.

2.3. Software and Libraries

The entire computational work of this research was implemented in Python (v. 3.8.18), in a Jupyter Lab environment. The libraries that were used for data preprocessing, the implementation of WGANs, and the data visualization were TensorFlow, Keras, Scikit-learn, SciPy, Pandas, NumPy, and Matplotlib.

2.4. Hyperparameter Tuning

The selection of the hyperparameters is of great importance in neural networks like WGANs. These hyperparameters include the activation function of the hidden layers and the output layers, the number of the hidden layers and the neurons of the generator and the critic, their learning rates, the epochs, the latent dimensions, and the batch size [31].

In this study, a variety of different combinations of the hyperparameters were examined to find the optimum set through which the WGANs work best, through the process of trial and error. Activation functions were tailored to input and output distributions; linear functions were used for one-dimensional input data in both the critic's hidden and output layers. Three different functions were explored for the generator output to match the critic input, with sigmoid and tanh functions tested to manage outliers caused by the linear function. Additionally, the clipping values, which refer to the range within which the weights of the critic's neural network are constrained during training, were $[-0.01, 0.01]$.

The hyperparameters explored before the application of WGAN in the Monte Carlo simulated clinical trials are displayed in Table 1. It should be underlined that after selecting the hyperparameter values, they were kept constant to ensure the reproducibility of the entire process.

Table 1. Hyperparameter tuning of the WGANs.

Activation Function		Number of Hidden Layers		Number of Neurons in Hidden Layers		Learning Rates	Epochs	Latent Dimensions	Batch Size
Hidden Layers	Output Layers	Generator	Critic	Generator	Critic				
ReLU	linear	1	1	12	12	0.01–0.0001	100	70	32
linear	sigmoid			8	8		150	80	64
	tanh			6	6		200	100	128
				4	4		250	120	256

2.5. Monte Carlo Simulations

In our research, we utilized a Monte Carlo simulation approach to generate random patient data for the evaluation of our algorithm. The Monte Carlo framework used in this study is the typical one that has been extensively used in simulated bioequivalence studies [10]. The reason for utilizing a Monte Carlo simulation framework in our study was to assess the performance of the WGAN algorithm across a range of simulated patient populations, capturing the variability inherent in real-world clinical scenarios. For the generation of the 10,000 “virtual patients”, two different normal distributions (T and R, referring to the Test and Reference treatment, respectively) were created. Each distribution had a mean endpoint value of 100 units, with the standard deviation varying according to the scenario explored. These normal distributions were created using the NumPy library. The function used to generate the distributions included three parameters: the mean of the distribution, the standard deviation, and the number of “patients”. Our goal was to mimic real clinical trial conditions, where an unknown population is sampled randomly to apply interventions in the sample drawn. The random sample maintained the format of the original population. Subsequently, statistical analysis is performed using the information gained from the sample. In our study, we simulated the population (referred to as the “original” dataset) from which we extracted the “sample”. We then progressed one step further by synthesizing virtual patients (i.e., the “generated” dataset). To obtain robust estimates, this process was repeated thousands of times using Monte Carlo simulations. It should be mentioned that in this manuscript the data source is the “original data” that come from simulations. However, in practical applications, the WGAN can easily process various formats of real-world clinical data (e.g., CSV, XLSX), making its implementation straightforward.

All data distributions were preprocessed through standard normalization (z-scores), and then a random sampling was performed. The sampling percentage was 0.5% and 1.0% of the population, which corresponds to 50 and 100 subjects, respectively. With that choice, the aim was to have a small representative “sample” of subjects from which the “generated” dataset would be created, using WGAN. This procedure was repeated for all scenarios listed in Table 2:

Table 2. Scenarios that were examined in the Monte Carlo simulations of the clinical trials.

Test/Reference Ratios	Coefficient of Variation for Between-Subject Variability	Size of the “Original” Dataset (Population)	Size of the “Generated” Dataset
1.0, 1.05, 1.1, 1.15	10%	10,000	10,000
1.0, 1.05, 1.1, 1.15	20%	10,000	10,000
1.0, 1.05, 1.1, 1.15	50%	10,000	10,000
1.0, 1.05, 1.1, 1.15	70%	10,000	10,000
1.0, 1.05, 1.1, 1.15	10%	10,000	1000
1.0, 1.05, 1.1, 1.15	10%	10,000	500
1.0, 1.05, 1.1, 1.15	10%	10,000	100

Four different ratios of the T/R ratio were examined where the mean value of the R group was kept constant to 100, while the T group had a mean of 100 multiplied by specific values: 1.0, 1.05, 1.1, and 1.15. The coefficient of variation (CV), for between-subject variability, was the same for both the T and R groups, determined by the scenario. The CV values indicate variability, with 10% representing low variability and 20% medium variability, while 50% and 70% refer to high and very high variability, respectively.

The statistical evaluation of the “original”, “sample”, and “generated” datasets was performed through the *t*-test criterion. A success in the *t*-test evaluation implied that the distributions (T and R) did not have any statistically significant differences.

The comparison of distributions across all scenarios was assessed using the Wasserstein distance, a metric that quantifies the dissimilarity between two probability distributions by measuring the minimum effort needed to transform one distribution into another. Minimizing this distance during WGAN training ensures that the generated dataset closely approximates the distribution of the original data.

Overall, the data generation and preprocessing procedures outlined above were integral components of our methodology for simulating synthetic clinical datasets. These steps were carefully designed to emulate real-world data characteristics while ensuring consistency and reliability in our experimental framework and will be highlighted in the revised manuscript.

3. Results

Initially, the scenarios studied referred to a 0.5% sampling rate, which corresponds to 50 patients (“sample”) from the initial 10,000 subjects in the population (“original”). From this low number of subjects, the WGAN approach was used to synthesize the “generated” dataset which was set to have equal size with the “original” (that is 10,000 generated subjects). The average histograms of two different scenarios can be seen in Figure 4. These histograms were calculated by creating individual histograms for each execution of the WGANs within a certain scenario and then averaging them. In Figure 4A, a scenario with a small T/R ratio and low variability was examined. A desired performance of the WGAN is observed towards reproducing a dataset (“generated”) with almost exactly identical distribution with the “original” data. Despite the fact that the “sample” did not appear to reflect the traits of the original population, due to the limited sample size, the generated dataset behaves similarly to the actual population. In Figure 4B, the results were consistent, even with a larger T/R ratio but the same variability. This indicates that the WGANs were able to capture the population performance from a sample that did not accurately represent the population.

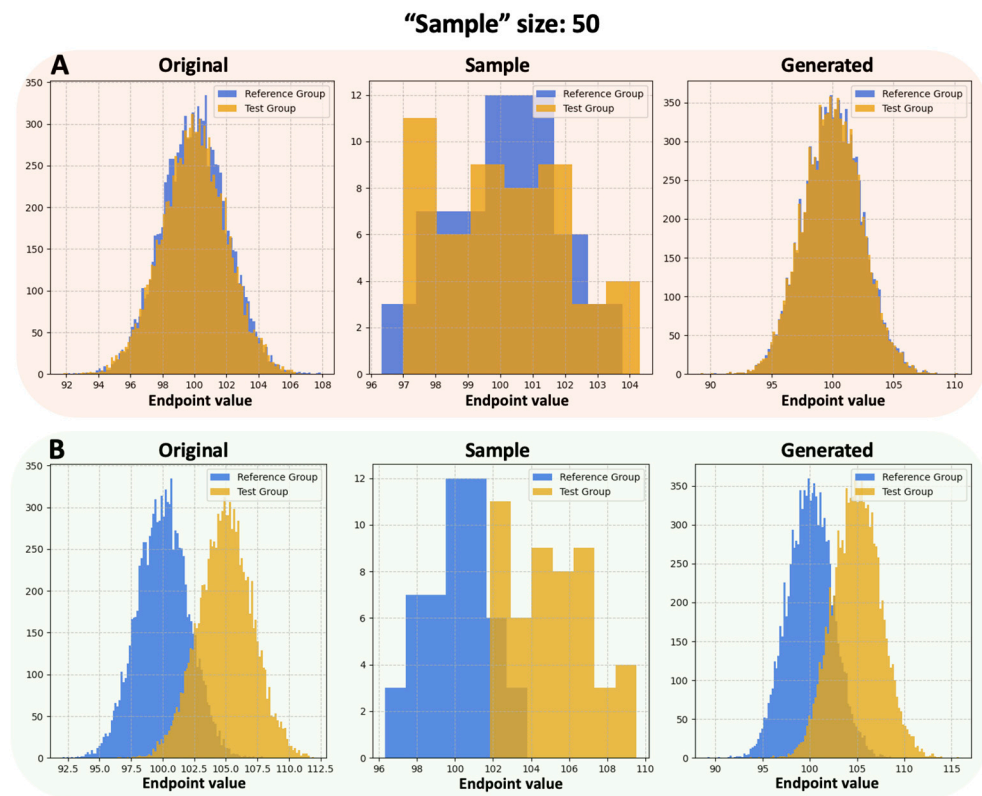


Figure 4. Distribution of the measured endpoints for the “original” (population), “sample” (50 subjects), and WGAN-“generated” datasets. The coefficient of variation for between-subject variability is 20%, while the Test/Reference ratio refers to 1.00 (A) and 1.05 (B).

Figure 4 shows that WGANs have the capability to synthesize virtual data with the same distribution shape. This aspect is crucial as it serves as a necessary prerequisite for the further assessment of performance among the three sets of data, namely, “original”, “sample”, and the WGAN-“generated”. Thus, the analysis could proceed to investigate the similarity in the statistical outcomes between the virtual population and the original population. In this context, the concept of “similarity”, as defined in Section 2.1, was assessed (Figure 5). It is reminded that similarity is assessed between two pairs: “original” vs. “generated” population, and “original” vs. “sample”.

Figure 5 clearly shows that across several different scenarios, the similarity between the “generated” population and the “original” population consistently exceeded that of the “original” vs. “sample”. These findings highlight the desired ability of the AI-synthesized data to mimic the behavior of the actual population. It also highlights the potential risk of false negative errors (i.e., type II error) within the traditional way we perform the clinical trials, a risk which is eliminated by virtually increasing the sample size with WGANs. Notably, the similarity percentage remains relatively stable for the “original” vs. “generated” population comparisons (minimum percentage: 94% and maximum percentage: 100%), indicating the successful imitation of population performance by the WGANs compared with the “original”–“sample pair” (minimum percentage: 4% and maximum percentage: 100%).

Apart from the similarities in the distributions, the success percentages were calculated and visualized, as illustrated in Figure 6 for the scenario of $a = T/R = 1.0$. The “generated” population exhibited higher or equal success percentages than the “original” and “sample” datasets across all different scenarios. Regarding the “sample”, with the increase in the T/R ratio, the successes were reduced. As expected, when the CV increases the successes of the “sample” increased. The number of successes for the “sample” were higher than the “original” and the “generated” population because of the small number of values. This also

supports the fact that the “sample” cannot reflect the “original” population, except when the variability is as low as 10% and the T/R ratio is equal to 1.0.

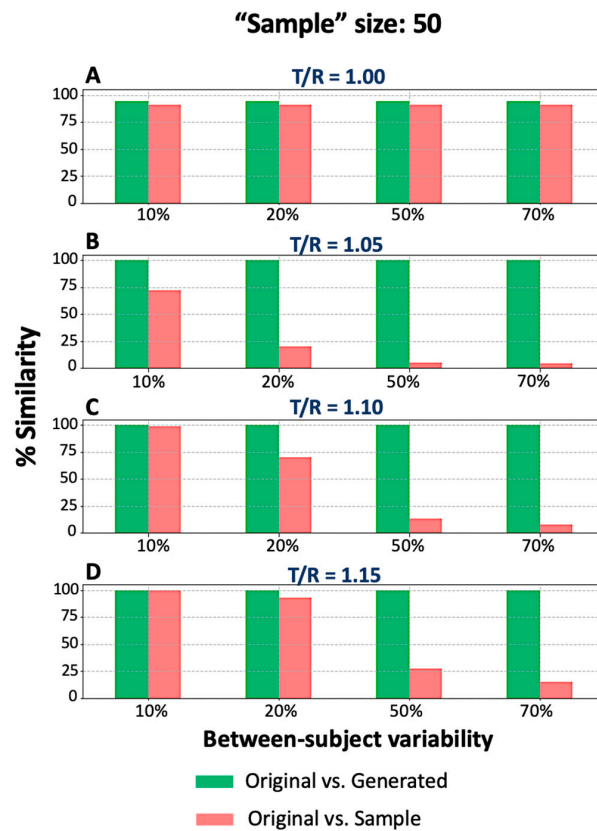


Figure 5. Percent similarity in the performance among the “original”, “sample”, and “generated” datasets when the size of the “sample” dataset is fixed at 50 subjects. The average Test/Reference (T/R) ratio is equal to 1.00 (A), 1.05 (B), 1.10 (C), and 1.15 (D). Four levels (10%, 20%, 50%, 70%) of between-subject variability were explored. Under each scenario, a number of 5000 Monte Carlo clinical trials are simulated.

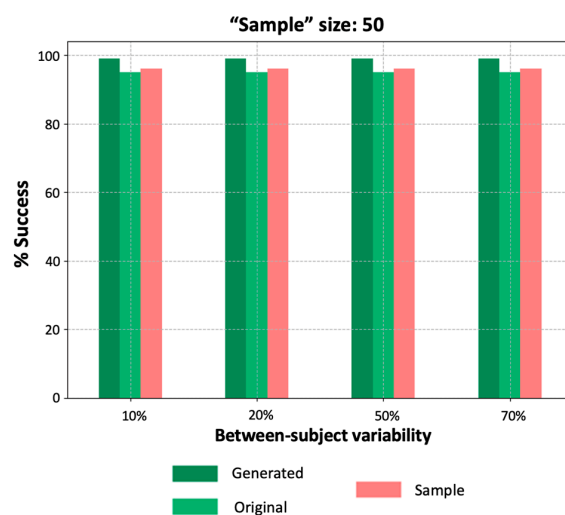


Figure 6. Percent of success for the “original”, “sample”, and “generated” datasets when the size of the “sample” dataset is fixed at 50 subjects. In all cases, the Test/Reference ratio is equal to unity. Four between-subject variability levels are used (10%, 20%, 50%, 70%). A success in the *t*-test evaluation means that the two treatments under comparison (Test vs. Reference) do not differ significantly (at the 5% level). Under each scenario, a number of 5000 Monte Carlo clinical trials are simulated.

After analyzing the results of the 0.5% sampling, a 1.0% sampling was conducted, representing 100 patients from the initial population of 10,000. Figure 7 provides a visual summary of the similarity percentages observed in this scenario.

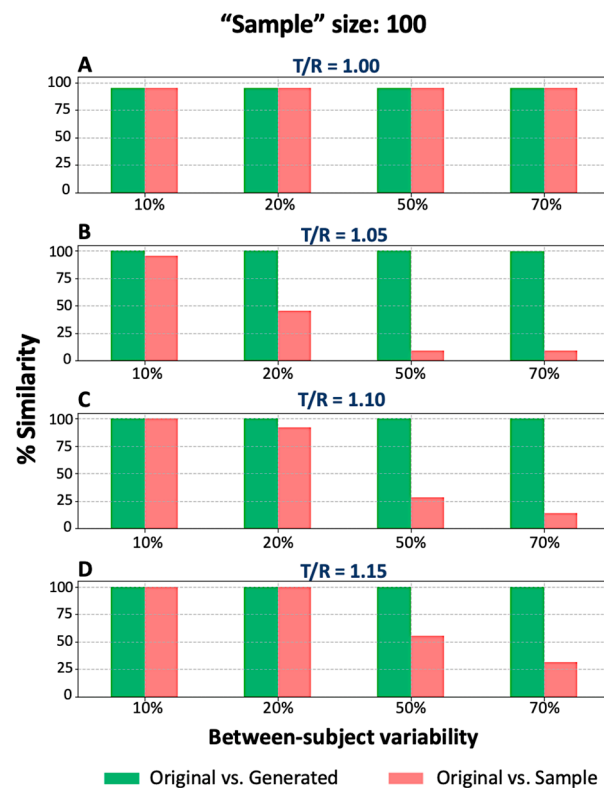


Figure 7. Percent similarity in the performance among the “original”, “sample”, and “generated” datasets when the size of the “sample” dataset is 100. The average Test/Reference (T/R) ratio is equal to 1.00 (A), 1.05 (B), 1.10 (C), and 1.15 (D). Four levels (10%, 20%, 50%, 70%) of between-subject variability are explored. Under each scenario, a number of 5000 Monte Carlo clinical trials are simulated.

From the 16 scenarios explored, the similarity percentages between the “original” and the “generated” population were greater than those between the “original” and the “sample” in 9 scenarios. For the remaining 7 scenarios, the similarities were equal. Notably, an increase in the T/R ratio corresponded to a decrease in similarity as the between-subject variability (in terms of CV) increased. Generally, WGANs exhibited a better handling of high variability compared with the sample, except when T/R equaled to 1.0, where the WGANs and the sample exhibited almost the same results.

As in the case of Figure 4, where the “sample” referred to 50 subjects, there is a clear similarity in the shape of distributions between the “original” and the “generated” datasets when the “sample” equals 100 subjects (Figure A1). Again, it is evident that the “generated” distributions closely resemble the “original”. It should be highlighted that the “sample” distributions failed to capture the “original” shape, whereas WGANs successfully generated a population that mirrored the original.

The success percentages for the 1.0% sampling (sample size = 100) were then visualized, in the case of T/R = 1 (Figure 8). It can be observed that the “generated” population demonstrated a higher success percentage than the “original”, while the “sample” exhibited a higher success percentage than both. Across all scenarios, the “generated” population showed either a higher or equal success percentage compared with the “original” population, while the “sample” had a higher success percentage in 13 out of 16 scenarios, with an equal success percentage observed in the remaining 3 scenarios compared with the “original” and “generated” population simultaneously. Even with a larger sample

percentage (twice the size of the one shown before), the “sample” exhibited an inferior behavior compared with the “original” or “generated” datasets in terms of similarity and statistical findings.

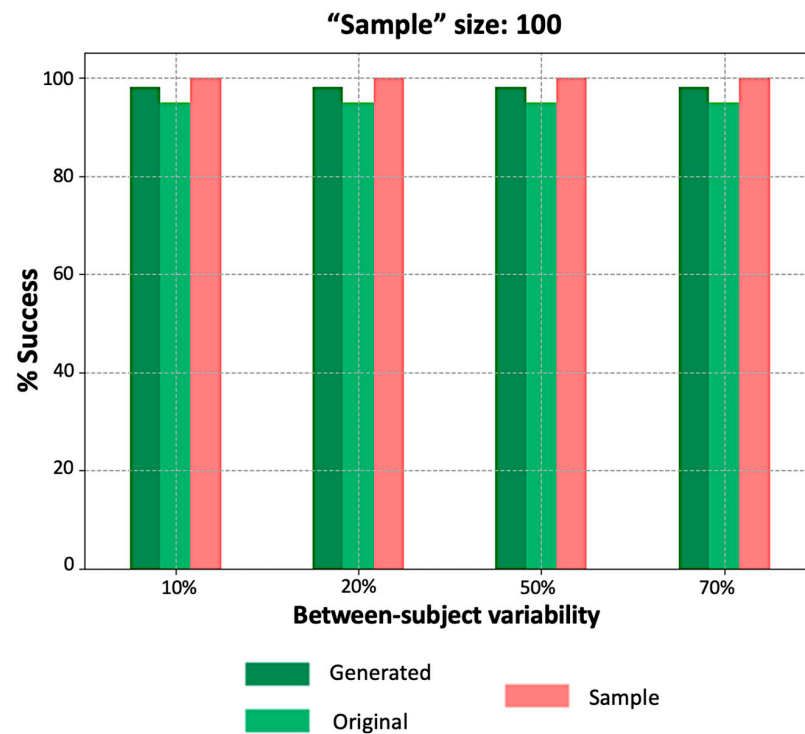


Figure 8. Percent of success for the “original”, “sample”, and “generated” datasets when the size of the “sample” dataset is fixed at 100 subjects. In all cases, the Test/Reference ratio is equal to unity. Four between-subject variability levels are used (10%, 20%, 50%, 70%). A success in the *t*-test evaluation means that the two treatments under comparison (Test vs. Reference) do not differ significantly (at the 5% level). Under each scenario, a number of 5000 Monte Carlo clinical trials are simulated.

In all previous scenarios, the size of the “generated” population was set to be equal to that of the “original” population, which was 10,000 subjects, and the observed performance of the WGAN procedure was found to be favorable. To further investigate the WGAN performance, smaller sizes were also explored for the “generated” datasets. In this context, Figure 9 presents the percent similarity in the performances of the three datasets when smaller sizes were used for the “generated” datasets, specifically referring to 1000, 500, and 100 subjects. In all these cases, the between-subject variability was set to 10%, the T/R ratio took four values (1.0, 1.05, 1.10, 1.15), and the size of the intermediate “sample” was either 50 or 100.

Across all three “generated” sizes (i.e., 1000, 500, 100) and two “sample” sizes (50 and 100), the similarity percentages were similar, particularly when the T/R ratio was 1.10 or 1.15. However, when the T/R ratio differed by 5%, from 1.0 to 1.05, the similarity between the “original” and “sample” decreased. It is noteworthy that the similarity percentage for the “generated” dataset compared with the “original” population consistently increased with each increment of the T/R ratio. The percent similarity reflects the success rate of the *t*-test evaluation, which is evident in the success percentages of the scenario with $\alpha = 1.0$ and $CV = 10\%$. These success rates are depicted in Figure A2 for all three sizes of the “generated” populations (i.e., 1000, 500, 100) and the two “sample” sizes (i.e., 50 and 100).

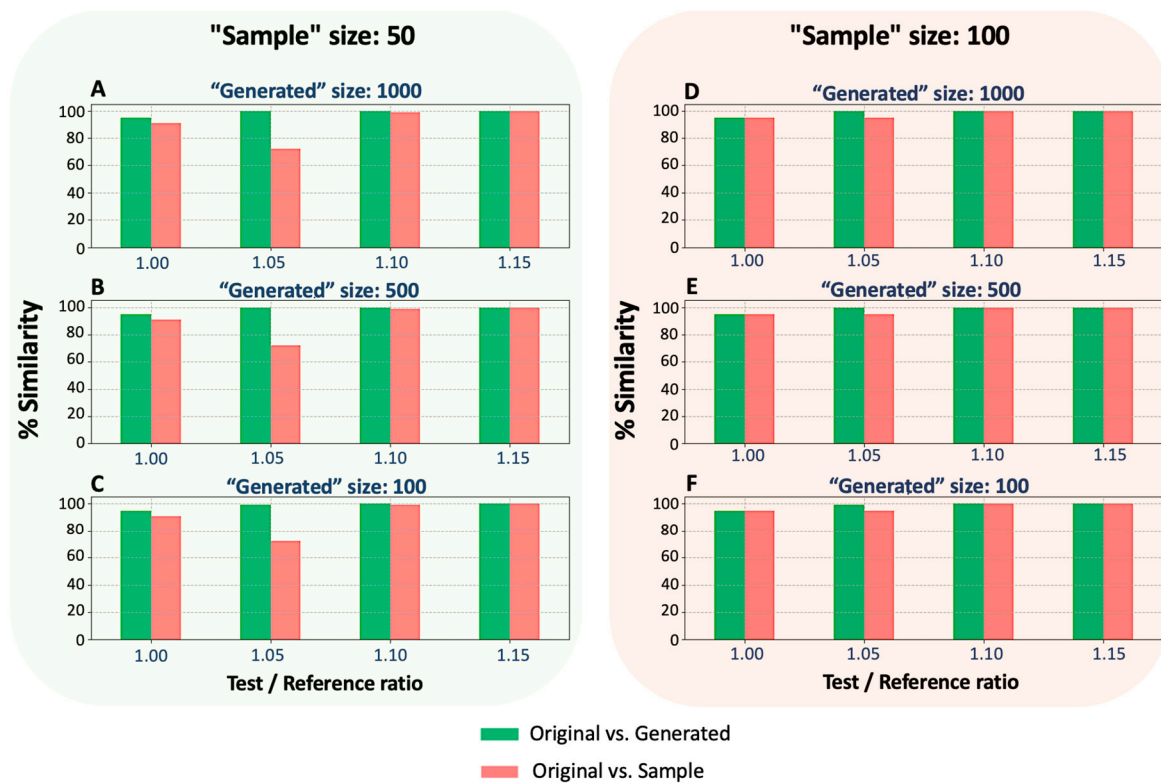


Figure 9. Percent similarity in the performance among the “original”, “sample”, and “generated” datasets when the size of the “sample” dataset is either equal to 50 (left panel) or 100 (right panel). The sizes of the “Generated” samples are equal to 1000 (A,D), 500 (B,E), and 100 (C,F). In all cases, the between-subject variability is 10%. Under each scenario, a number of 5000 Monte Carlo clinical trials are simulated.

Wasserstein distances were computed between the “original” and “generated” populations for both Test and Reference groups after each iteration. On average, these distances closely mirrored the product of the T/R ratio and the coefficient of variation. The Wasserstein distance, along with the similarity parameter, was considered to evaluate the degree of resemblance between the two distributions.

In evaluating the performance of the WGAN and its susceptibility to variations in input data quality and diversity, our study maintained consistent WGAN hyperparameters across diverse scenarios. Notably, despite the introduction of variability in datasets containing varying patient numbers (ranging from 10,000 to 100), our findings indicate that the WGAN functioned effectively without necessitating a recalibration of the algorithmic parameters. This underscores the robustness of the WGAN architecture in accommodating diverse input data profiles, thereby showcasing its adaptability and reliability across a spectrum of clinical study settings.

4. Discussion

The aim of this study is to propose a novel approach for data augmentation in clinical research by utilizing generative AI algorithms like WGANs. Bentley et al. [32] have highlighted persistent delays in clinical trials, impacting both budgets and research outcomes. Inadequate participant recruitment leads to low-value trials and resource wastage, while the involvement of the private sector in trial financing adds complexity to the landscape. Furthermore, trial duration emerges as a significant cost driver, with each additional month in phase III trials resulting in a median expenditure of \$671,000. Consequently, even marginal reductions in trial cycle times could yield substantial cost savings in overall clinical development budgets [33].

The exponential emergence of AI methodologies during the last years started to grow within the pharmaceutical and healthcare sectors. The utilization of AI-driven algorithms enables healthcare professionals to analyze extensive patient datasets, uncovering patterns that may elude human clinicians initially. This capability facilitates early disease detection, prompt medical interventions, and improved patient prognoses. Additionally, AI has the potential to predict treatments most likely to yield positive outcomes for individual patients, enabling personalized medicine tailored to unique patient characteristics. By optimizing administrative processes and reducing redundant tests, AI also holds promise for reducing healthcare expenses, both in terms of finances and time [11].

In this context, this study introduces a novel method aimed at decreasing sample size in clinical trials by using WGANs. The proposed idea is simple: to draw conclusions from a small sample in a clinical trial yet achieve the same statistical power as that of an entire population in a clinical trial. As achieving this using traditional statistical methods is impossible, we used a generative AI algorithm (WGANs) to tackle this challenge.

The newly proposed method involves training WGANs on a limited sample and utilizing it to generate “virtual patients”. To evaluate the efficacy of the current method, we conducted Monte Carlo simulations of clinical studies to compare the performance of the WGAN-generated virtual population (referred to as the “generated” dataset) with both the original population and a small sample. The proposed approach suggests that by combining a small sample from the population with a generative algorithm like WGANs, it is possible to create a virtual population. This can enhance statistical power without raising type I error rates.

To accomplish this task, Monte Carlo simulations were applied to generate a virtual population of 10,000 patients, representing the entire population, referred to as the “original” dataset. To create the “sample”, we mimic the process used in real-world scenarios by randomly sampling from the “original” dataset. In this study, two small proportions of the “original” dataset were utilized (0.5% and 1.0%) corresponding to 50 and 100 subjects in the “sample”, respectively. WGANs were then used to exploit the information contained in the “sample” and to synthesize the “generated” dataset (Figure 2). Initially, the size of the “generated” data was set as equal to the population size, i.e., 10,000 subjects. Subsequently, smaller sizes of the “generated” dataset were used in the analysis, such as 1000, 500, and 100. Pairwise comparisons were then made among the “original”, “sample”, and “generated” datasets to assess their performance. Also, similarity with the “original” dataset (i.e., the population) was estimated in order to express the concordance in the statistical results between the WGAN-“generated” dataset and the “sample”. The performance of the WGANs towards the scope of the study was assessed under several scenarios like different between-subject variability, difference between the two treatments (expressed as Test/Reference ratios), “sample” size, “generated” size, etc.

In summary, the application of WGANs demonstrated favorable performance. Across all scenarios examined, the AI-generated dataset resulted in better results compared with the “sample” and, in most cases, the performance of WGANs was found to be significantly superior to that of the traditional sample. Starting with the 0.5% sampling (equivalent to 50 patients from the population), the “sample” fails to capture the population traits (Figure 4). Conversely, the “generated” population, shown in Figure 5, consistently mirrors the population characteristics, indicating successful imitation by WGANs. These findings highlight the desired ability of the AI-synthesized data to mimic the behavior of the actual population. The similarity between the “generated” and “original” populations consistently surpassed that of the “original”–“sample” pair across scenarios, except in one scenario ($T/R = 1.15$, $CV = 10\%$). It also highlights the potential risk of false negative errors (i.e., type II error) within the traditional way we perform the clinical trials, a risk which is eliminated by virtually increasing the sample size with WGANs. Conversely, as the T/R ratio increases, success rates decrease for the sample, while they rise with higher variabilities (i.e., CV). Also, with the 1.0% sampling, representing 100 patients from the initial 10,000, across 16 scenarios, the similarity percentages between the “original” and

the “generated” populations were higher in nine scenarios, with seven scenarios showing equal similarities. Notably, as the T/R ratio increased, similarity decreased with higher CVs. WGANs generally handled high variability better than the “sample”, except when the T/R was 1.0, where both showed similar results (Figures 7 and A1). Also, in all scenarios explored, the “generated” population consistently showed higher or equal success rates compared with the “original” data (i.e., the population). In the above-mentioned scenarios, the size of the “generated” population was matched to that of the “original” population, comprising 10,000 subjects, and the observed performance of the WGAN procedure was deemed rather satisfactory. To obtain deeper insights into the WGAN performance, smaller sizes were also examined for the “generated” datasets (Figure 9). Again, in these cases, the similarity percentages for the “generated” sizes of 1000, 500, and 100 were similar across both sampling percentages. Also, the success rates were similar across all “generated” sizes, T/R ratios, and samplings, with generated population success matching that of the population. To this point, it is important to underline some key points about the proposed method. Initially, using synthetic data in clinical trials aims only to improve the statistical power of the study. This is achieved by utilizing WGANs to create virtual subjects based on the limited sample size. This work suggests using WGANs not to replace human volunteers entirely but rather partially. The aim is to reduce human involvement in clinical trials, thus boosting statistical power without needing more participants, thus avoiding ethical concerns, costs, and time limitations associated with recruiting additional human participants. However, it is essential to have a sufficient and representative sample of human volunteers to accurately generate virtual subjects. And it is also worth mentioning that data synthesized by WGANs are not intended to be used for safety evaluation. Another crucial point is reproducibility [34–36]. Ensuring consistency in the model architecture, hyperparameters, and random seed makes WGANs reproducible. We conducted in-depth hyperparameter tuning to optimize the performance of WGANs. Various parameter combinations were tested to evaluate the algorithm. It is also vital to prevent hallucinations of generated data. In this study, we took all necessary measures to ensure the entire process is fully reproducible [37]. Nevertheless, the regulatory authorities should establish specific criteria and guidelines for the application of AI-generated virtual subjects in practice [38–40]. Finally, the importance of data augmentation has been recognized for many years. In this context, techniques like bootstrapping have routinely been used to increase sample sizes in clinical trials, but they have limitations such as overfitting, bias, loss of information, and assuming independence [41]. These restrictions are due to the fact that the same subjects are used, unlike the synthesis of new data as in the case of generative AI algorithms (e.g., WGANs).

Notably, GANs find extensive applications across diverse domains. For example, in medical imaging, they are instrumental in generating synthetic MRI images for brain tumor detection [42]. In biomedical signal processing, GANs enhance the detection of epileptic activity through signal augmentation, while in computer vision, they can perform tasks like speech emotion recognition [43,44]. GANs also play a pivotal role in data augmentation, improving machine learning model performance across domains like medical imaging for autism detection [45]. Recently, GANs have been used for security and privacy applications, generating synthetic data to anonymize sensitive information, creating adversarial examples for evaluating model robustness, and synthesizing realistic images for training surveillance systems [46]. It is noteworthy that GANs contribute to entertainment and creativity, generating artistic content including artwork, music, and literature, thereby fostering new avenues for artistic expression and creativity [47].

The need for rigorous validation and comprehensive ethical review of the virtual data generated by any generative AI algorithm should be highlighted. The validation process, among others, should include the following steps: comparison with real-world data, performance benchmarks, and external validation. For the first step, it is necessary to conduct extensive validation studies by comparing the characteristics of the synthetic data with actual clinical trial data to ensure fidelity and representativeness. This will involve

statistical analyses to assess distribution, variance, and other key metrics. For the second step, the generative model should be benchmarked against established datasets, so its performance can be evaluated using industry-standard metrics to confirm its accuracy and reliability. The third step will involve collaboration with independent researchers to perform external validation of the synthetic datasets, ensuring objectivity and transparency in the validation process [48].

Furthermore, ethical review must be approved by the Institutional Review Board (IRB). Prior to any implementation in clinical settings, the generated data and the processes involved should undergo thorough review and approval by an IRB or an equivalent ethics committee. There must also be compliance with regulations to ensure that the computational methods comply with all relevant guidelines, such as the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), to protect patient privacy and data security. Finally, detailed documentation of the data generation process, including the algorithms used and the steps taken to validate and secure the data, should be maintained, and made available for review by ethical committees and regulatory bodies [49].

Ethical considerations for clinical studies should involve obtaining consent, where applicable, for the use of clinical data in generating synthetic datasets from patients or their representatives. Therefore, appropriate measures should be taken to identify and reduce any biases in the synthetic data generation process to ensure fairness and representativeness of the synthetic datasets. By following these steps, it is anticipated that the synthetic data generated by a generative AI algorithm are both scientifically valid and ethically sound for use in clinical research [50].

The study has some limitations that need to be mentioned. Firstly, the time duration of each scenario to complete all iterations poses a limitation, potentially affecting the robustness of the results. While the precise timeframes and resource allocations may vary depending on the specific experimental conditions, we can provide an approximate range based on our experiments. Typically, model training ranged from 20 to 30 min, with larger datasets and more complex architectures requiring longer training times than the smaller ones. Furthermore, we recognize the importance of optimizing computational efficiency and scalability in future iterations of our methodology. However, in real practice, simulations will not be required; WGANs would be directly applied to the actual “sample” human data. In our study, model training was conducted using relatively modest computational resources. Specifically, we utilized two computers equipped with the Windows 11 operating system, each powered by an Intel(R) Core(TM) i5-9400 CPU at 2.90 GHz and equipped with 8 GB of installed RAM. Although these systems are not as powerful as dedicated high-performance computing clusters, they still provided sufficient computational capability to successfully train our model. The Monte Carlo simulations conducted in this study were solely performed to evaluate the applicability of the WGAN approach compared with the traditional “sample”-based procedure. Another limitation is the absence of real clinical data in this study; incorporating real-world clinical data will enhance the accuracy and reliability of WGANs, enabling better predictions and informed decision-making in clinical studies.

It is important to clarify that our initial use of virtual generated data served as a foundational step in exploring the feasibility and potential benefits of data augmentation techniques, such as the implementation of the WGAN, within the context of clinical studies. As such, our focus is primarily centered on evaluating the efficacy and performance of the WGAN algorithm in generating synthetic data that closely resemble real-world clinical datasets. Training the algorithm with real-world problems would be particularly beneficial in addressing these limitations and improving the utility of WGANs in clinical research. Another issue that should be mentioned relies on the fact that only the simple *t*-test criterion was used. While the primary focus was to evaluate the efficacy of the proposed algorithm in generating synthetic patient data across various scenarios, there is potential for further investigation. For instance, including an additional random error following a normal

distribution as a reference point for comparison, and comparing WGANs with other methodologies or algorithms, would enrich the analysis. The WGAN data augmentation approach should be further explored in various types of clinical trials, including non-inferiority, superiority, and bioequivalence studies, as well as implementing real-life data from clinical trials to assess the feasibility of the algorithm.

Overall, the integration of synthetic data into various research fields is already underway. To our knowledge, there has been no prior exploration of implementing the WGAN algorithm to generate “virtual patients” aimed at reducing the time and expenses associated with clinical studies and minimizing human exposure. Two recent studies [17,18] by our research group, highlighted the positive role of variational autoencoders in clinical and bioequivalence trials. This study expands these two previous works by using an alternative generative AI algorithm, the Wasserstein Generative Adversarial Networks.

5. Conclusions

In this study, a new approach aimed at reducing the requirement for large sample sizes in clinical trials is proposed. The new method involves training an artificial neural network, specifically a WGAN, on a limited dataset (the typical “sample” of clinical study) and then applying the WGAN to generate virtual subjects. To demonstrate the suitability of the WGAN-based approach, a new methodological procedure was also required to be established and applied. Monte Carlo simulations of clinical studies were conducted to compare the performance of the WGAN-generated virtual subjects (the “generated” dataset) against both the entire population (referred to as the “original” dataset) and a subset of it, known as the “sample”. The percent successes, as well as the similarity in the performance compared with the “original” population, were recorded. It was shown that only a small subset (the “sample”) of the true population can be utilized alongside WGANs to create a virtual population (the “generated” dataset), and the latter can lead to similar results as the entire population. Overall, this work aims to provide a potential solution to the challenges associated with small sample sizes and human exposure in clinical research.

Author Contributions: Conceptualization, V.D.K.; methodology, V.D.K.; software, A.N.; validation, A.N.; formal analysis, A.N.; investigation, A.N.; resources, V.D.K.; data curation, A.N.; writing—original draft preparation, A.N.; writing—review and editing, V.D.K.; visualization, A.N. and V.D.K.; supervision, V.D.K.; project administration, V.D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

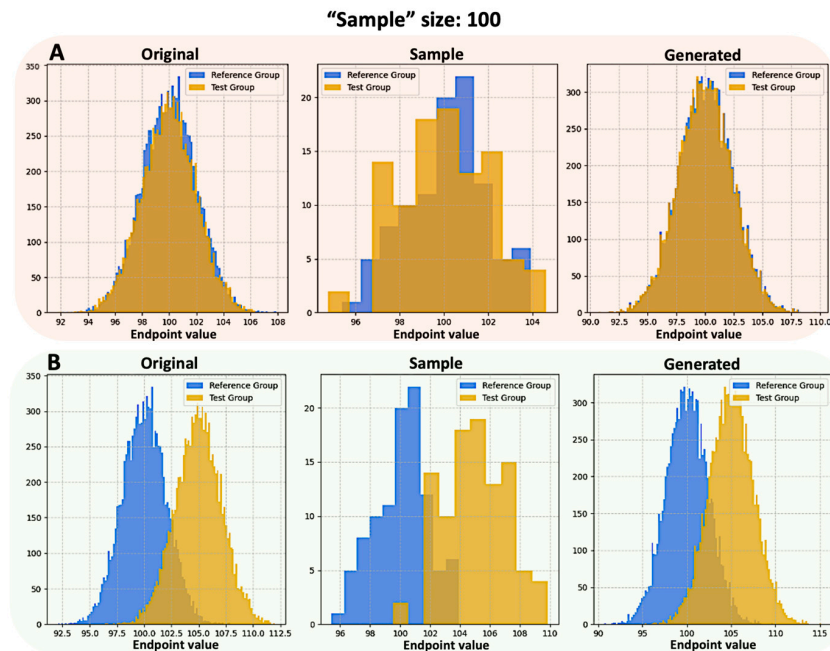


Figure A1. Distribution of the measured endpoints for the “original” (population), “sample” (100 subjects), and WGAN-“generated” datasets. The coefficient of variation for between-subject variability was 20%, while the Test/Reference ratio refers to 1.00 (A) and 1.05 (B).

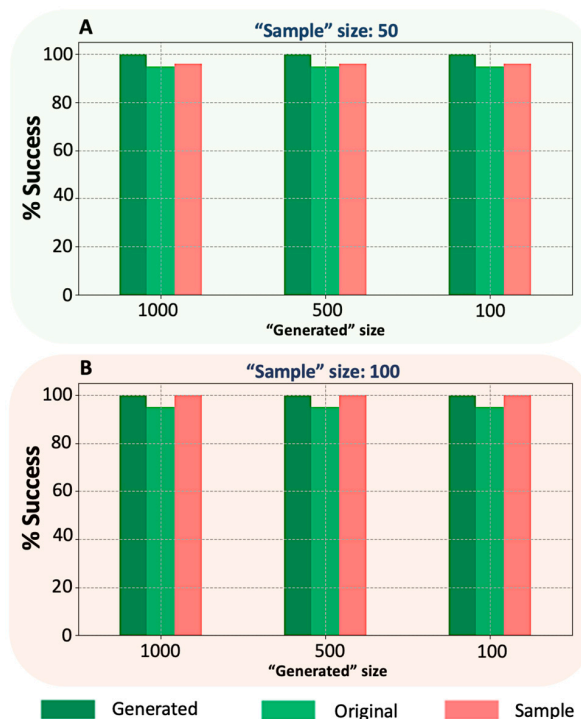


Figure A2. Percent of success for the “original”, “sample”, and “generated” datasets when the size of the “sample” dataset is either equal to 50 (A), or 100 (B). Three different values for the size of the “generated” dataset are used: 1000, 500, and 100. In all cases, the Test/Reference ratio is equal to unity, while the between-subject variability level is 10%. A success in the *t*-test evaluation means that the two treatments under comparison (Test vs. Reference) do not differ significantly (at the 5% level). Under each scenario, a number of 5000 Monte Carlo clinical trials are simulated.

References

1. Wang, X.; Ji, X. Sample Size Estimation in Clinical Research: From Randomized Controlled Trials to Observational Studies. *Chest* **2020**, *158*, S12–S20. [[CrossRef](#)] [[PubMed](#)]
2. Sakpal, T.V. Sample Size Estimation in Clinical Trial. *Perspect. Clin. Res.* **2010**, *1*, 67–69. [[CrossRef](#)] [[PubMed](#)]
3. Andrade, C. Sample Size and Its Importance in Research. *Indian J. Psychol. Med.* **2020**, *42*, 102–103. [[CrossRef](#)] [[PubMed](#)]
4. Serdar, C.C.; Cihan, M.; Yücel, D.; Serdar, M.A. Sample Size, Power and Effect Size Revisited: Simplified and Practical Approaches in Pre-Clinical, Clinical and Laboratory Studies. *Biochem. Medica* **2021**, *31*, 010502. [[CrossRef](#)] [[PubMed](#)]
5. Ji, Z.; Lin, J.; Lin, J. Optimal Sample Size Determination for Single-Arm Trials in Pediatric and Rare Populations with Bayesian Borrowing. *J. Biopharm. Stat.* **2022**, *32*, 529–546. [[CrossRef](#)] [[PubMed](#)]
6. Hajian-Tilaki, K. Sample Size Estimation in Diagnostic Test Studies of Biomedical Informatics. *J. Biomed. Inform.* **2014**, *48*, 193–204. [[CrossRef](#)] [[PubMed](#)]
7. Brookes, S.T.; Whitely, E.; Egger, M.; Smith, G.D.; Mulheran, P.A.; Peters, T.J. Subgroup Analyses in Randomized Trials: Risks of Subgroup-Specific Analyses. *J. Clin. Epidemiol.* **2004**, *57*, 229–236. [[CrossRef](#)] [[PubMed](#)]
8. Martin-McGill, K.J.; Bresnahan, R.; Levy, R.G.; Cooper, P.N. Ketogenic Diets for Drug-Resistant Epilepsy. *Cochrane Libr.* **2020**, *2020*, CD001903. [[CrossRef](#)] [[PubMed](#)]
9. Wang, S.S.; Canida, T.A.; Ihrie, J.D.; Chirtel, S.J. Sample Size Determination for Food Sampling. *J. Food Prot.* **2023**, *86*, 100134. [[CrossRef](#)]
10. Karalis, V. Modeling and Simulation in Bioequivalence. In *Interdisciplinary Applied Mathematics*; Springer International Publishing: Cham, Switzerland, 2016; pp. 227–254.
11. Karalis, V.D. The Integration of Artificial Intelligence into Clinical Practice. *Appl. Biosci.* **2024**, *3*, 14–44. [[CrossRef](#)]
12. Gupta, R.; Srivastava, D.; Sahu, M.; Tiwari, S.; Ambasta, R.K.; Kumar, P. Artificial Intelligence to Deep Learning: Machine Intelligence Approach for Drug Discovery. *Mol. Divers.* **2021**, *25*, 1315–1360. [[CrossRef](#)] [[PubMed](#)]
13. Ramesh, A.N.; Kambhampati, C.; Monson, J.R.T.; Drew, P.J. Artificial Intelligence in Medicine. *Ann. R. Coll. Surg. Engl.* **2004**, *86*, 334–338. [[CrossRef](#)] [[PubMed](#)]
14. Ossowska, A.; Kusiak, A.; Świątlik, D. Artificial Intelligence in Dentistry—Narrative Review. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3449. [[CrossRef](#)]
15. Hashimoto, D.A.; Witkowski, E.; Gao, L.; Meireles, O.; Rosman, G. Artificial Intelligence in Anesthesiology Current Techniques, Clinical Applications, and Limitations. *Anesthesiology* **2020**, *132*, 379–394. [[CrossRef](#)] [[PubMed](#)]
16. Keskinbora, K.; Güven, F. Artificial Intelligence and Ophthalmology. *Türk Oftalmol. Derg.* **2020**, *50*, 37–43. [[CrossRef](#)] [[PubMed](#)]
17. Papadopoulos, D.N.; Karalis, V. Variational Autoencoders for Data Augmentation in Clinical Studies. *Appl. Sci.* **2023**, *13*, 8793. [[CrossRef](#)]
18. Papadopoulos, D.; Karalis, V.D. Introducing an Artificial Neural Network for Virtually Increasing the Sample Size of Bioequivalence Studies. *Appl. Sci.* **2024**, *14*, 2970. [[CrossRef](#)]
19. Maharana, K.; Mondal, S.; Nemade, B. A Review: Data Pre-Processing and Data Augmentation Techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [[CrossRef](#)]
20. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. GAN (Generative Adversarial Nets). *J. Jpn. Soc. Fuzzy Theory Intell. Inform.* **2017**, *29*, 177.
21. Ahmad, W.; Ali, H.; Shah, Z.; Azmat, S. A New Generative Adversarial Network for Medical Images Super Resolution. *Sci. Rep.* **2022**, *12*, 9533. [[CrossRef](#)]
22. Zhang, F.; Wang, L.; Zhao, J.; Zhang, X. Medical Applications of Generative Adversarial Network: A Visualization Analysis. *Acta Radiol.* **2023**, *64*, 2757–2767. [[CrossRef](#)] [[PubMed](#)]
23. Paladugu, P.S.; Ong, J.; Nelson, N.; Kamran, S.A.; Waisberg, E.; Zaman, N.; Kumar, R.; Dias, R.D.; Lee, A.G.; Tavakkoli, A. Generative Adversarial Networks in Medicine: Important Considerations for This Emerging Innovation in Artificial Intelligence. *Ann. Biomed. Eng.* **2023**, *51*, 2130–2142. [[CrossRef](#)] [[PubMed](#)]
24. Tanaka, F.H.K.D.S.; Aranha, C. Data Augmentation Using GANs. *arXiv* **2019**, arXiv:1904.09135.
25. Wang, W.; Pai, T. Enhancing Small Tabular Clinical Trial Dataset through Hybrid Data Augmentation: Combining SMOTE and WCGAN-GP. *Data* **2023**, *8*, 135. [[CrossRef](#)]
26. Patil, M.; Patil, M.M.; Agrawal, S. WGAN for Data Augmentation. In *GANs for Data Augmentation in Healthcare*; Springer International Publishing: Cham, Switzerland, 2023; pp. 223–241.
27. Das, S.; Dey, R.; Nayak, A.K. Artificial Intelligence in Pharmacy. *Indian J. Pharm. Educ. Res.* **2021**, *55*, 304–318. [[CrossRef](#)]
28. Mehmood, A.; Iqbal, M.; Mehmood, Z.; Irtaza, A.; Nawaz, M.; Nazir, T.; Masood, M. Prediction of Heart Disease Using Deep Convolutional Neural Networks. *Arab. J. Sci. Eng.* **2021**, *46*, 3409–3422. [[CrossRef](#)]
29. Chow, S.; Shao, J.; Wang, H.; Lokhnygina, Y. *Sample Size Calculations in Clinical Research*, 3rd ed.; Informa UK Limited: London, UK, 2017.
30. Krenmayr, L.; Frank, R.; Drobig, C.; Braungart, M.; Seidel, J.; Schaudt, D.; von Schwerin, R.; Stucke-Straub, K. GANerAid: Realistic Synthetic Patient Data for Clinical Trials. *Inform. Med. Unlocked* **2022**, *35*, 101118. [[CrossRef](#)]
31. Chollet, F. *Deep Learning with Python*; Simon and Schuster: New York, NY, USA, 2021.
32. Bentley, C.; Cressman, S.; van der Hoek, K.; Arts, K.; Dancey, J.; Peacock, S. Conducting Clinical Trials—Costs, Impacts, and the Value of Clinical Trials Networks: A Scoping Review. *Clin. Trials* **2019**, *16*, 183–193. [[CrossRef](#)]

33. Martin, L.; Hutchens, M.; Hawkins, C.; Radnov, A. How Much Do Clinical Trials Cost? *Nat. Rev. Drug Discov.* **2017**, *16*, 381–382. [[CrossRef](#)]
34. Foster, D. *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*; O'Reilly & Associates Inc.: Sebastopol, CA, USA, 2019.
35. Liu, C.; Gao, C.; Xia, X.; Lo, D.; Grundy, J.; Yang, X. On the Reproducibility and Replicability of Deep Learning in Software Engineering. *ACM Trans. Softw. Eng. Methodol.* **2022**, *31*, 1–46. [[CrossRef](#)]
36. Chien, J.-T. Deep Neural Network. In *Source Separation and Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 259–320.
37. Verma, S.; Tran, K.; Ali, Y.; Min, G. Reducing LLM Hallucinations Using Epistemic Neural Networks. *arXiv* **2023**, arXiv:2312.15576.
38. Dykstra, K.; Mehrotra, N.; Tornøe, C.W.; Kastrissios, H.; Patel, B.; Al-Huniti, N.; Jadhav, P.; Wang, Y.; Byon, W. Reporting Guidelines for Population Pharmacokinetic Analyses. *J. Pharmacokinet. Pharmacodyn.* **2015**, *42*, 301–314. [[CrossRef](#)] [[PubMed](#)]
39. FDA; Population Pharmacokinetics Guidance for Industry; U.S. Department of Health and Human Services Food and Drug Administration; Center for Drug Evaluation and Research (CDER); Center for Biologics Evaluation and Research (CBER). 2022. Available online: <https://www.fda.gov/media/128793/download> (accessed on 14 April 2024).
40. EMA. Guideline on Reporting the Results of Population Pharmacokinetic Analyses. Committee for Medicinal Products for Human Use (CHMP). 2007. Available online: <https://www.ema.europa.eu/en/reporting-results-population-pharmacokinetic-analyses-scientific-guideline> (accessed on 14 April 2024).
41. Klinger, C. *Bootstrapping Reality from the Limitations of Logic: Developing the Foundations of “Process Physics”, a Radical In-Formation-Theoretic Modelling of Reality* Paperback-22; VDM Publishing: Riga, Latvia, 2010.
42. Xu, Z.; Qi, C.; Xu, G. Semi-supervised attention-guided CycleGAN for data augmentation on medical images. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; IEEE: Piscataway, NJ, USA, 2019. [[CrossRef](#)]
43. Wei, Z.; Zou, J.; Zhang, J.; Xu, J. Automatic epileptic EEG detection using convolutional neural network with improvements in time-domain. *Biomed. Signal Process. Control* **2019**, *53*, 101551. [[CrossRef](#)]
44. Shilandari, A.; Marvi, H.; Khosravi, H.; Wang, W. Speech emotion recognition using data augmentation method by cycle-generative Adversarial Networks. *Signal Image Video Process.* **2022**, *16*, 1955–1962. [[CrossRef](#)]
45. Bouallegue, G.; Djemal, R. EEG data augmentation using Wasserstein Gan. In Proceedings of the 2020 20th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), Monastir, Tunisia, 20–22 December 2020; IEEE: Piscataway, NJ, USA, 2020. [[CrossRef](#)]
46. Cai, Z.; Xiong, Z.; Xu, H.; Wang, P.; Li, W.; Pan, Y. Generative Adversarial Networks: A Survey Toward Private and Secure Applications. *ACM Comput. Surv.* **2021**, *54*, 132. [[CrossRef](#)]
47. Shahriar, S. GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. *Display* **2022**, 102237. [[CrossRef](#)]
48. Wu, Y.; Kumar, A. *Machine Learning and Artificial Intelligence in Healthcare Systems*; CRC Press: Boca Raton, FL, USA, 2020.
49. Steyerberg, E.W. *Clinical Prediction Models*; Springer: Berlin/Heidelberg, Germany, 2019.
50. Barocas, S.; Hardt, M.; Narayanan, A. *Fairness and Machine Learning*. 2019. Available online: <https://fairmlbook.org/pdf/fairmlbook.pdf> (accessed on 14 April 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.