

Article

Bioequivalence Studies of Highly Variable Drugs: An Old Problem Addressed by Artificial Neural Networks

Dimitris Papadopoulos ¹, Georgia Karali ^{2,3} and Vangelis D. Karalis ^{1,3,*} ¹ Department of Pharmacy, School of Health Sciences, National and Kapodistrian University of Athens, 15784 Athens, Greece; dimitrios.papadopoulos@pharm.uoa.gr² Department of Mathematics and Applied Mathematics, University of Crete, 71003 Heraklion, Greece³ 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: Bioequivalence studies of highly variable drugs require the utilization of large numbers of volunteers. The EMA and FDA propose the utilization of scaled limits. In this study, we introduce the use of artificial neural networks, along with the typical 80–125% limits, as a tool for virtually increasing sample size and thus reducing the actual human exposure.

Abstract: The bioequivalence (BE) of highly variable drugs is a complex issue in the pharmaceutical industry. The impact of this variability can significantly affect the required sample size and statistical power. In order to address this issue, the EMA and FDA propose the utilization of scaled limits. This study suggests the use of generative artificial intelligence (AI) algorithms, particularly variational autoencoders (VAEs), to virtually increase sample size and therefore reduce the need for actual human subjects in the BE studies of highly variable drugs. The primary aim of this study was to show the capability of using VAEs with constant acceptance limits (80–125%) and small sample sizes to achieve high statistical power. Monte Carlo simulations, incorporating two levels of stochasticity (between-subject and within-subject), were used to synthesize the virtual population. Various scenarios focusing on high variabilities were simulated. The performance of the VAE-generated datasets was compared to the official approaches imposed by the FDA and EMA, using either the constant 80–125% limits or scaled BE limits. To demonstrate the ability of AI generative algorithms to create virtual populations, no scaling was applied to the VAE-generated datasets, only to the actual data of the comparators. Across all scenarios, the VAE-generated datasets demonstrated superior performance compared to scaled or unscaled BE approaches, even with less than half of the typically required sample size. Overall, this study proposes the use of VAEs as a method to reduce the necessity of recruiting large numbers of subjects in BE studies.

Keywords: bioequivalence; highly variable drugs; artificial neural networks; variational autoencoders; stochasticity; Monte Carlo simulations



Citation: Papadopoulos, D.; Karali, G.; Karalis, V.D. Bioequivalence Studies of Highly Variable Drugs: An Old Problem Addressed by Artificial Neural Networks. *Appl. Sci.* **2024**, *14*, 5279. <https://doi.org/10.3390/app14125279>

Academic Editor: Burkhard Poeggeler

Received: 27 May 2024

Revised: 13 June 2024

Accepted: 17 June 2024

Published: 18 June 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

Bioequivalence (BE) testing aims to determine whether two drug products containing the same active ingredient are equivalent when administered in vivo [1,2]. Specifically, it compares a test drug (T) to an innovator's formulation, known as the reference product (R). The core of BE assessment lies in comparing the pharmacokinetic properties of the two drug products. This involves a detailed statistical analysis, including calculating a 90% confidence interval (CI). BE is typically declared if this 90% CI falls within the established range of 80–125% [1,2].

While this standard method, known as average BE, is widely accepted, it is not suitable for highly variable drugs or drug products. The expression “highly variable drugs” refers

to those with a within-subject coefficient of variation of 30% or more, whether due to the drug substance or its formulation [3]. In BE studies, this variability refers to the residual variability that comes from the ANOVA analysis after excluding all other known factors. In the case of a 2×2 crossover design, residual variability is estimated after subtracting from the total data variability the variability attributed to subjects, periods, sequences, and the administered pharmaceutical product.

Variability can arise from factors such as the drug characteristics or physiological conditions in patients. However, as within-subject variability increases, demonstrating BE becomes more challenging without increasing the sample size [3]. To address this issue, various methods have been proposed. Both the EMA and the U.S. FDA currently recommend using scaled BE limits [1,2]. This approach adjusts the BE limits based on the within-subject variability of the reference product. In this context, the EMA and FDA's recommended reference-scaled procedures require full-replicate or semi-replicate study designs. In these designs, the reference product is administered at least twice to each subject, allowing for accurate estimation of within-subject variability [1,2].

This study proposes the integration of generative artificial intelligence (AI) algorithms, specifically variational autoencoders (VAEs), to virtually increase sample sizes in BE studies of highly variable drugs. By using VAEs, which are a type of neural network designed to generate new data samples that mimic the original dataset, synthetic data can be created that closely resemble real patient data. This approach could significantly reduce the number of human subjects required for BE studies, addressing the challenges posed by high within-subject variability. It should be mentioned that sample size in BE testing is defined as the total number of patients or individuals that take part in the study [4]. Estimating the optimal number of participants is not always straightforward, considering that there are many different methods and formulas described in the literature and textbooks and in many cases a lot of inputs are required, like the expected effect size and the type I and II errors [5–9]. The advantages of an optimal sample size in studies are evident, such as an increase in statistical power, a reduction in bias, and an improvement in precision, accuracy, robustness, and the ability to conduct subgroup analyses [10–15].

By augmenting the dataset virtually, VAEs can increase the statistical power of the study, making it easier to demonstrate BE without the logistical and ethical complexities associated with large clinical trials. This innovative use of AI has the potential to streamline the BE assessment process, reduce costs, and expedite the availability of generic medications to the market. 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 [16–21]. Data augmentation has not been widely used in the healthcare industry; however a recent study from our lab has shown the efficiency and advantages of this approach in the field of clinical trials [22].

The aim of this study is to address the problem of the BE assessment of highly variable drugs by using artificial neural networks, particularly VAEs. This work relies on the inherent advantage of VAEs to act as noise filters, i.e., to reduce the variability of the original data, and builds on this to exploit their use in BE. Several scenarios with high variability, as well as other aspects of BE studies, are explored. In addition to the aforementioned scenarios, we investigated the performance of VAEs in comparison with the scaled BE limits imposed by the FDA and EMA [1,2].

2. Materials and Methods

2.1. General

The methodology framework of our study relies on two main components: VAEs and Monte Carlo simulations of BE studies. These components were used to demonstrate the applicability and advantages of VAEs in the context of BE studies of highly variable drugs. Specifically, BE trials were simulated under various conditions of within-subject variability, sample size, and differences between the T and R products [3]. Using Monte

Carlo simulations, two populations (termed the “original” dataset) of subjects were created for both the T and R groups across two periods, which were then randomly subsampled with different proportions (“subsampled”). Subsequently, the subsampled dataset was utilized to train a VAE model, and that trained VAE model was then used to generate new datasets (“generated”).

The conditions for creating the original data varied in terms of sample size, within-subject variability, and the mean ratio of the T and R endpoints (T/R). The VAE model generated datasets of varying sizes relative to the original sample size (1 to 3 times the original sample size). Additionally, the (unscaled) VAE-generated datasets were compared against the original and the subsampled datasets with the scaled BE approach in order to achieve a stricter evaluation of the performance of the VAE method (Figure 1).

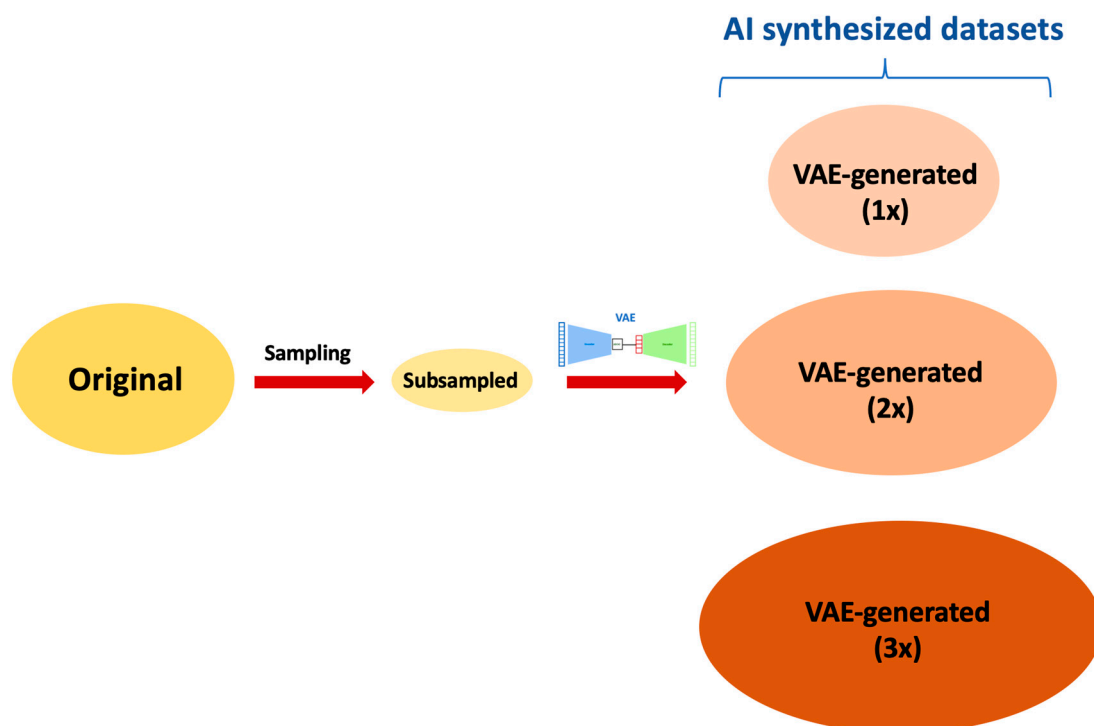


Figure 1. A graphical illustration of the general idea for using VAEs as a tool for data augmentation in bioequivalence studies with high-variability drugs. Instead of requiring a large sample size (termed “original”), only a subgroup of this (i.e., “subsampled”) is needed. VAE is then applied to the subsampled dataset in order to synthesize the generated datasets. The latter can exhibit either the same size (1×), double the size (2×), or triple the size (3×) of the original dataset.

BE assessment was applied to all three datasets (original, subsampled, and generated), adjusting the acceptance limits appropriately based on within-subject variability as per the guidelines of the EMA and the FDA [1,2]. This procedure was repeated thousands of times (5000) to ensure robust estimates. The total of 5000 Monte Carlo trials is based on findings from prior research [3]. These studies have demonstrated that conducting this quantity of repetitions ensures the acquisition of reliable and consistent estimates [3]. Each step of the analysis is described below.

2.2. Neural Networks and VAEs

Neural networks are algorithms that mimic the human brain structure and operation [23]. They consist of layers of stacked neurons that are interconnected with synapses [24,25]. A vector $x = [x_1, x_2, x_3, \dots, x_n]$ is inputted to the neural network. The input vector is multiplied with a vector of weights $w_i^j = [w_{i1}^j, w_{i2}^j, w_{i3}^j, \dots, w_{in}^j]$ and by adding a bias term (i.e., b_i^j). The linearly combined weight vector and input vector are then transformed

by a function (i.e., f) and are later used as input by the i th neuron at the j th layer. This procedure is described in Equation (1) and is usually referred to as “forward propagation”.

$$z_i = f(w_i^j \cdot x + b_i^j) \quad (1)$$

where $\cdot$ denotes the dot (scalar) product and f is called the activation function. The most common activation functions are the sigmoid function (2), the hyperbolic tangent (3), the rectified linear unit (i.e., ReLU) (4), and softplus (5):

$$f(z) = \frac{1}{1 + e^{-z}} \quad (2)$$

$$f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (3)$$

$$f(z) = \max(0, z) \quad (4)$$

$$f(z) = \ln(1 + e^z) \quad (5)$$

This process is repeated across all layers and neurons in a left-to-right manner. The final output of the network $\hat{y}$ is compared with the actual value y and the weights are eventually iteratively optimized with respect to a cost function that is suitable to the problem (backward propagation) [26]. The forward propagation calculations for the first neuron of the first layer are visually presented in Figure 2.

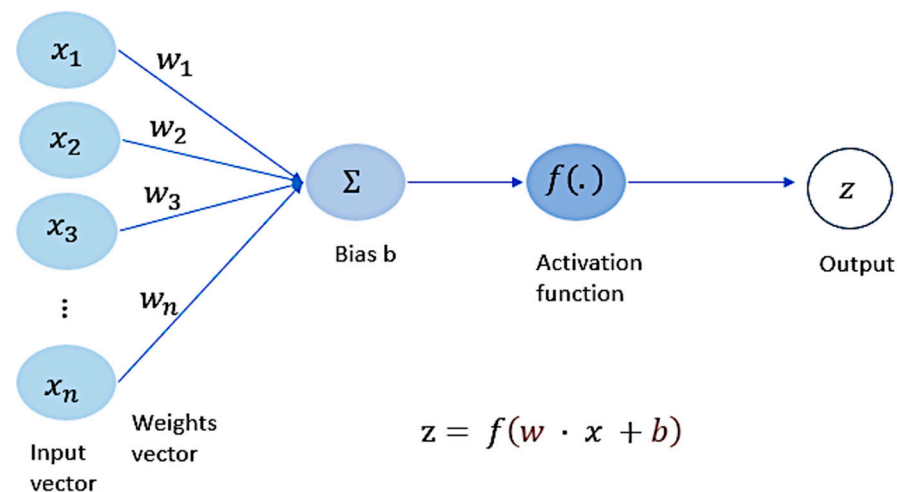


Figure 2. Visual representation of the calculations for the first neuron of the first layer during forward propagation.

VAEs are a special case of neural networks that can be used to generate new data based on existing data. They are composed of two parts: the “encoder” and its “mirrored” image, the decoder [22,27]. The encoder and the decoder are linked with the “latent space”. The input data are passed through the encoder and are mapped to a probability distribution in the latent space [28,29]. Datapoints from the latent space are then passed through the decoder, leading to the output. The differences between the traditional autoencoder and the VAE is that the loss function of the latter is composed of two parts, the reconstruction loss and the Kullback–Leibler loss. Moreover, random sampling can be performed from the probability distribution in the latent space. The sampled datapoints can be used as input from the decoder to generate novel datapoints. An extensive description of VAEs and autoencoders was presented in our previous studies [22,27].

A more in-depth explanation of variational autoencoders is provided in Appendix A and Figure A1.

2.3. Technical Aspects

Training a neural network requires as input a set of hyperparameters. This includes the number of layers, the activation function, the number of neurons at each layer, and the number of epochs. Choosing the optimal set of hyperparameters highly affects the performance of the model, and it vastly depends on the complexity and nature of the problem [30]. The tuning of the hyperparameters was achieved via trial and error, as well as by using information in the literature. In this study, we experimented with an extensive number of configurations, which are presented in Table 1. After testing all these tested combinations, it was found that the activation functions that worked best for the hidden and output layers were softplus and linear, respectively, while the optimal number of hidden layers was 3. In addition, the number of neurons that was found to be the best for each of the hidden layers, from left to right, was 64–32–16 for the encoder and 16–32–64 for the decoder. The ratio of the training set to the validation set was 4:1.

Table 1. List of hyperparameters used for training the VAE model. The latent space dimension and the number of epochs were set to 1 and 1000, respectively.

Number of Neurons in the Hidden Layers		Number of Hidden Layers		Activation Function	
Encoder	Decoder	Encoder	Decoder	Hidden Layers	Output Layer
128–64–32– 16–8–4	4–8–16–32– 64–128	2–5	2–5	Softplus	Softplus
				ReLU	Softmax
				ELU	Sigmoid
				Linear	Linear
				Tanh	Tanh

The process of adjusting the biases and weights of a network is referred to as back-propagation [26,27]. The primary goal during this phase is to minimize the cost function. In our specific application, the cost function was defined as a typical loss function where the Kullback–Leibler loss component and the reconstruction loss were equally weighted. The gradient of the error function with respect to the weights was used to update them incrementally during training, following a backward propagation approach aimed at minimizing the error function. This iterative process, involving both forward and backward propagation across the entire dataset, is termed an “epoch” in machine learning. At the end of each epoch, the loss function, which represents the error, is computed. The objective is to determine an appropriate number of epochs such that the error converges. The number of epochs required may vary depending on the complexity of the problem. In this study, the convergence of the algorithm was evaluated based on the stability of the loss function value computed at the conclusion of each epoch. This indicates that the value of the loss function stabilized by the final epoch. Based on plots of the loss function versus the number of epochs (Figure A2) and after exploring the similarity of the input/output distribution (Figure A3), the number 1000 was considered the optimal choice.

The entire computational work was performed in Python version 3.7 with TensorFlow 2.10.0. In our investigation, we used modest computational resources for training. Specifically, we used two computers (Fujitsu Siemens, Athens, Greece) running the Windows 11 operating system, each powered by an Intel® Core™ i5-9400 CPU at 2.90 GHz with 8 GB of RAM. While these systems are not as robust as dedicated high-performance computing clusters, they still offered adequate computational power to effectively train our model. Generally, model training typically lasted between 20 and 30 min, with larger datasets and more intricate architectures necessitating longer training durations compared to smaller ones.

2.4. Simulation of Bioequivalence Studies

In the case of a typical crossover design for BE studies, two treatments were administered: the reference drug (i.e., R) and the test drug (i.e., T), across two periods [1–3]. Half of the subjects received the R drug in period 1 and the T drug in period 2 (sequence RT), while the other half received the T drug in period 1 and the R drug in period 2 (sequence TR). After study completion, a specific statistical framework was applied, and the BE of the two drugs was either accepted or rejected based on the comparison between the 90% confidence interval and the acceptance limits. Generally, for highly variable drugs, the acceptance limits for the BE test are scaled according to the within-subject variability. Various regulatory authorities, such as the EMA and FDA, permit different acceptance limits for highly variable pharmaceutical products [1,2].

The first step of the framework was the random generation of N subjects for both groups for the first period—specifically, N_R and N_T individuals for the R and T groups, respectively, with mean μ_{R1} and standard deviation σ_{R1} for the R group and mean μ_{T1} and standard deviation σ_{T1} for the T group. Both groups had equal CV (i.e., the stochastic term for between-subject variability) and sample size; thus, $N_R = N_T$ and $N = N_R + N_T$. Later, the N individuals were multiplied with a randomly generated “stochastic term,” with mean 1 and standard deviation σ_w . This stochastic term (with coefficient of variation CV_W) represented the within-subject variability [3]. It is important to note that the CV_W is different than the between-subject variability (i.e., CV) that was mentioned before.

Afterwards, the aforementioned N individuals for both periods that are termed “original” were subsampled with proportions of 50%, 75%, and 100% (termed “subsampled”). The subsampled subjects were used to train the hyperparameter-tuned VAE model to generate new subjects, termed “generated”. Finally, BE testing was conducted between the T and R groups across the original, subsampled, and generated datasets by utilizing the standard statistical criteria mandated by the EMA and FDA depending on the value of CV_W , and the success or failure of the statistical test was recorded. This procedure was repeated 5000 times in order to obtain robust estimates for the % BE acceptance. A list of factors analyzed in the study, including the CV_W , the relationship between μ_{T1} and μ_{R1} , N , the subsampled proportions, and the size of the generated datasets proportionally to the total size N , are listed in Table 2.

Table 2. Factors relevant to the simulated bioequivalence studies explored in this study. For all cases, the mean of the stochastic term expressing within-subject variability was set to 1.

Original Sample Size (N)	Within-Subject Variability (CV_W)	Ratio of Average Endpoints Test/Reference	Between-Subject Variability (CV)	Size of Generated Dataset (xN)
12	20%	1	20%	1×
24	40%	1:1		2×
48	60%			3×
72				

3. Results

The first step of this study was to explore the BE acceptance rates for all three types of datasets (i.e., original, subsampled, and generated) under the condition that both the T and R drugs have equal means in the measured endpoint.

Figure 3 illustrates the BE acceptance rates for different coefficients of variation of 20% (Figure 3A), 40% (Figure 3B), and 60% (Figure 3C). For the highly variable cases (40% and 60%), the scaled BE limits, imposed by the EMA or FDA, were used in the case of the original and subsampled datasets in accordance with the regulatory guidelines [3].

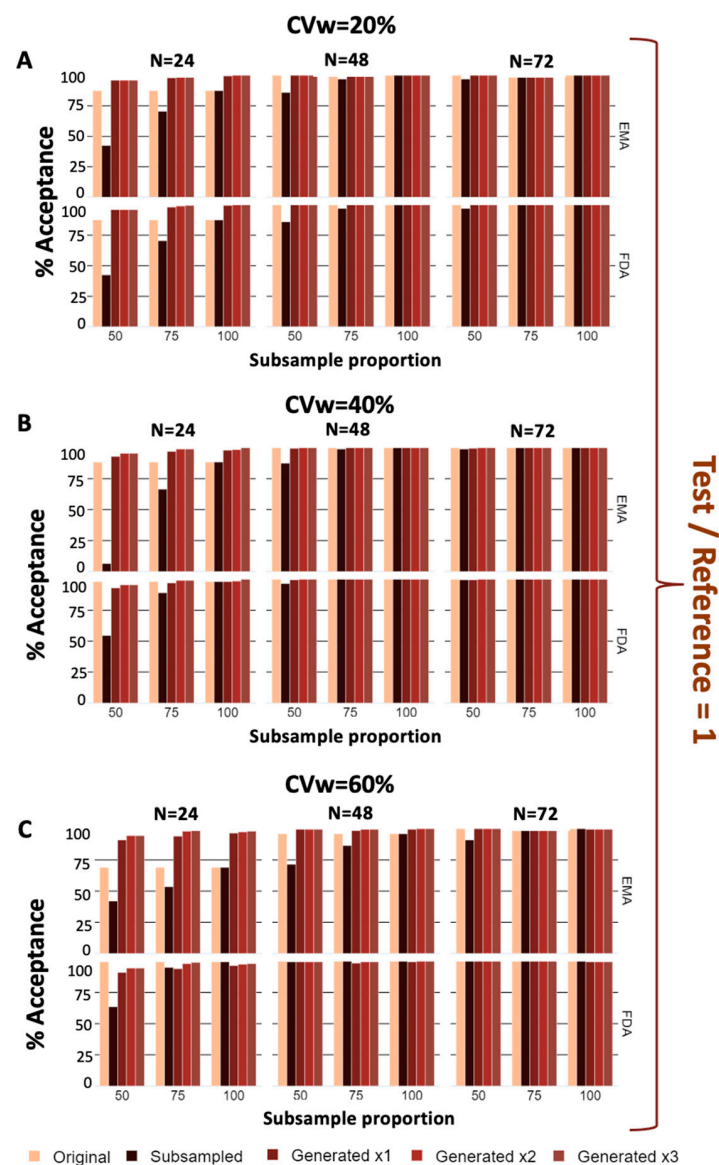


Figure 3. Acceptance rates of the original, subsampled, and generated datasets. In all cases, the average test/reference ratio is 1 and the within-subject variability equals 20% (A), 40% (B), and 60% (C). Three different original sample sizes are used (24, 48, and 72), while the sizes of the generated datasets are 1, 2, and 3 times that of the original sample. The bioequivalence assessment for the original and subsampled dataset is performed using the EMA and FDA scaled bioequivalence approaches [1,2], while no scaling is applied to the VAE-generated datasets. The subsample proportions used are 50%, 75%, and 100%.

No scaling was used for the VAE-generated datasets, in order to be treated more strictly. Aiming to assess the statistical power achieved, the BE acceptance rates were explored for various sample and subsample sizes. Figure 3 shows that for all datasets, the acceptance rate decreases as the coefficient of variation increases, while the power increases for larger original and subsample sizes. It is shown that the percentage of acceptance of the subsampled set is always lower than that of the original, whereas the acceptance rate for the generated is always higher than that of the original dataset. It is important to emphasize that for 40% and 60% variabilities, the acceptance limits when conducting BE testing for the generated datasets, did not follow the scaled BE approach, while the original and the subsampled datasets were extended as a function of CV_W , in line with the EMA guidelines.

A similar behavior is observed for the same scenarios under the assumption that the ratio of the means of the T and R pharmaceutical products is 1:1, which is displayed in Figure 4. There is an increasing trend of the acceptance rate for all types of datasets when the sample and subsample size increases for both the FDA and the EMA. The original datasets outperform the subsampled datasets in all cases, whereas the VAE-generated datasets outperform the original datasets for CV = 20% (Figure 4A) and CV = 40% (Figure 4B). It is important to underline that for the scenarios presented in Figure 4, the acceptance limits for the VAE-generated datasets were not extended, in contrast with the acceptance limits for the original and subsampled cases.

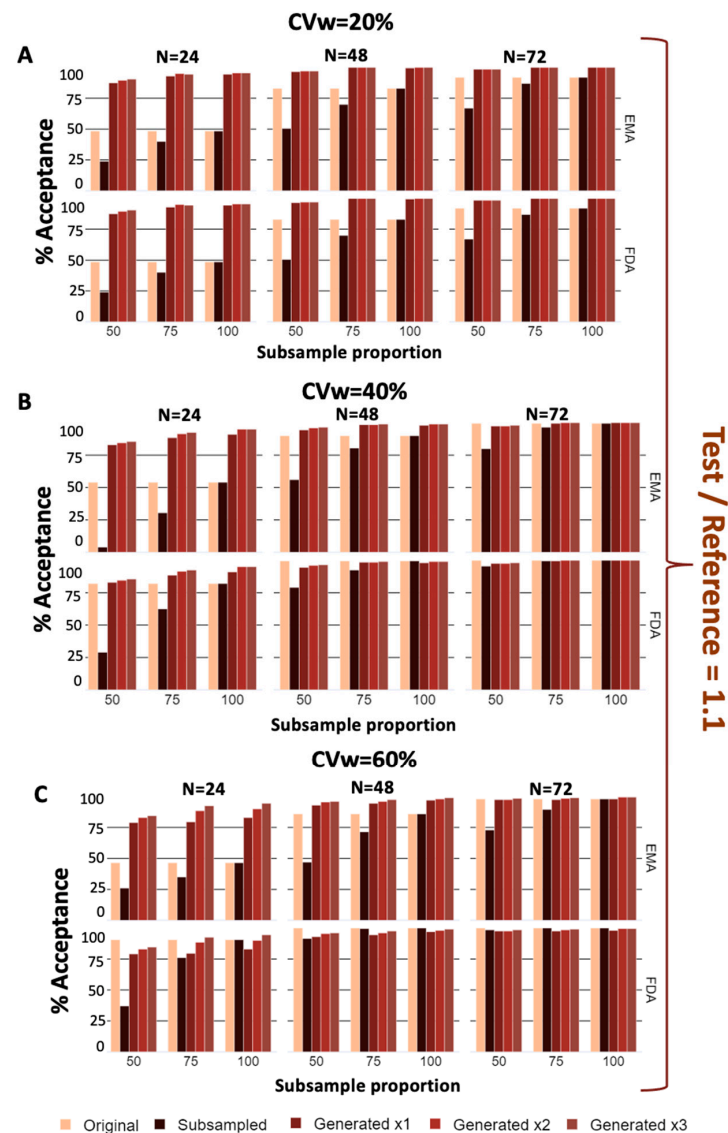


Figure 4. Acceptance rates of the original, subsampled, and generated datasets. In all cases, the average test/reference ratio is 1:1 (i.e., 10% mean difference) and the within-subject variability equals 20% (A), 40% (B), and 60% (C). Three different original sample sizes are used (24, 48, and 72), while the sizes of the generated datasets are 1, 2, and 3 times that of the original sample. The bioequivalence assessment for the original and subsampled datasets is performed using the EMA and FDA scaled bioequivalence approaches [1,2], while no scaling is applied to the VAE-generated datasets. The subsample proportions used are 50%, 75%, and 100%.

To obtain a clearer comparison between the performance of the VAE approach and the classic/scaled limits, the percentage of “acceptance gains” of BE were calculated for each

scenario studied. These acceptance gains were defined as the acceptance rate of the VAE-generated dataset minus the acceptance rate of the original dataset. The acceptance gains were calculated for all the scenarios listed in Table 2 and are presented in Figures 5 and 6. In Figure 5, the ratio of the means of the T and R pharmaceutical products is equal to 1, while in Figure 6 the T/R ratio is 1:1. The scenarios that are not shown in Figures 5 and 6 refer to the cases where there was no acceptance gain, namely, both the original and the generated datasets exhibited similar performance in terms of the statistical power.

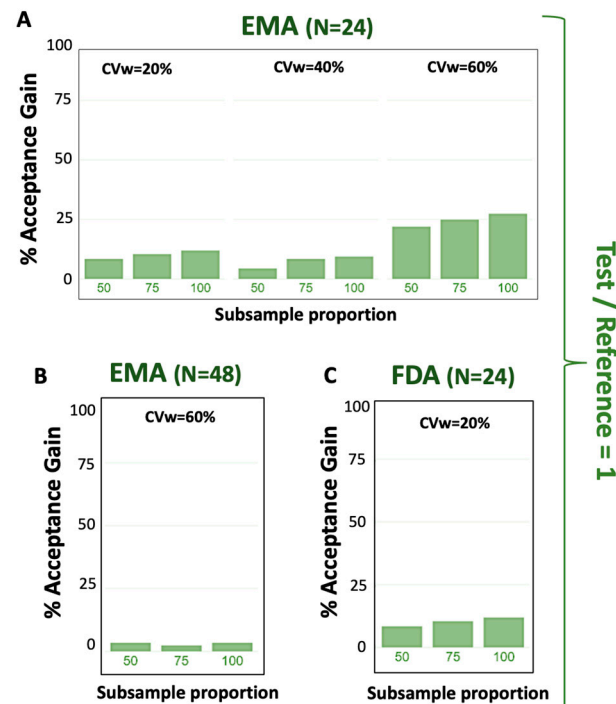


Figure 5. Acceptance gains, namely, the difference in the acceptance rate percentages between the VAE-generated and the original datasets. The acceptance limits for the original datasets are increased according to the EMA (A,B) and FDA (C) guidelines. In all cases, the average test/reference ratio is 1 and the within-subject variability equals 20%, 40%, and 60%. Three different original sample sizes are used (24, 48, and 72), while the size of the generated datasets is equal ($1 \times$) to that of the original sample. The scenarios where the generated and original datasets perform identically are not shown due to space restrictions.

Figure 5 illustrates that the acceptance gain for the EMA scenarios (Figures 5A and 5B for 24 and 48 subjects, respectively) is achieved when the CV and sample size are equal to 60% and 24, respectively. For the FDA case (Figure 5C), the acceptance gain ranges from 10% to 15% for the scenario where the CV equals 20% and the sample size equals 24. For both regulatory authorities, the acceptance gain increases overall as the subsample size increases.

Figure 6 presents the acceptance gain for the scenario where T/R equals 1:1. Figure 6A, 6B and 6C show the acceptance gain when the limits are extended according to the EMA for sample sizes of 24, 48, and 72, respectively. Figure 6D–F show the acceptance gain when the limits are extended according to the FDA for the pre-mentioned sample sizes. In the case of the EMA (Figure 6A–C), the acceptance gain ranges from 5% when the CV and sample size refer to 40% and 48%, respectively, to 48% for the scenario where the CV equals 20% and the sample size equals 24%. The acceptance gain range in the FDA cases (Figure 6D–F) is similar to that of the EMA, though there are more scenarios where no acceptance gain was observed.

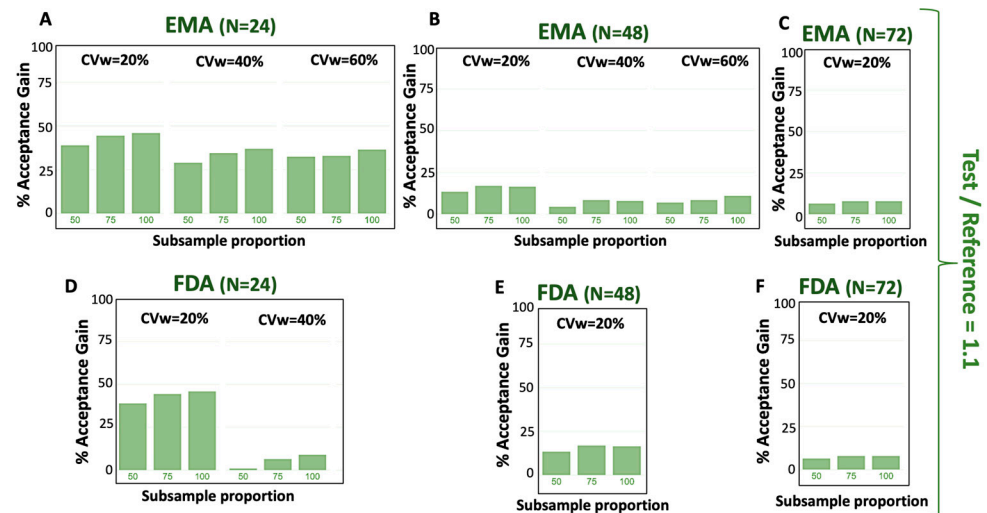


Figure 6. Acceptance gains, namely, the difference in the acceptance rate percentages between the VAE-generated and the original datasets. The acceptance limits for the original datasets are increased according to the EMA (A–C) and FDA (D–F) guidelines. In all cases, the average test/reference ratio is 1:1 and the within-subject variability equals 20%, 40%, and 60%. Three different original sample sizes are used (24, 48, and 72), while the size of the generated datasets is equal ($1\times$) to that of the original sample. The scenarios where the generated and original datasets perform identically are not shown due to space restrictions.

Figures 7 and 8 show the percentage of acceptance gain when no scaling is applied to the original dataset. Therefore, this refers to a condition where both the VAE-generated and the original datasets are treated equally. Scenarios for T/R equal to 1 and 1:1 are shown, along with sample sizes of 24, 48, and 72.

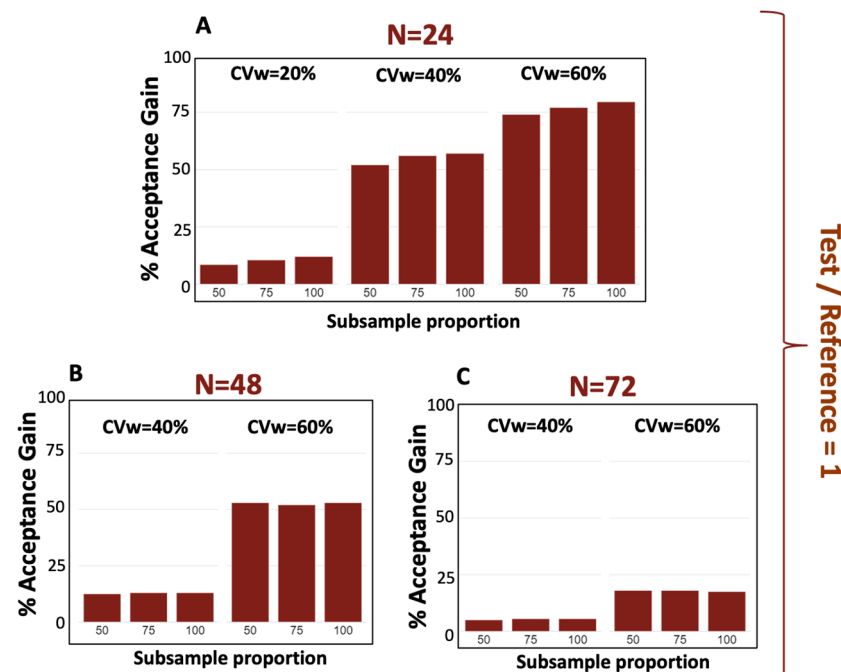


Figure 7. Acceptance gains between the VAE-generated and the original datasets, when no scaling is applied to either dataset. Three different original sample sizes are used (24 (A), 48 (B), and 72 (C)), while the size of the generated datasets is equal ($1\times$) to that of the original sample. In all cases, the average test/reference ratio is 1 and the within-subject variability equals 20%, 40%, and 60%. The scenarios where the generated and original datasets perform identically are not shown due to space restrictions.

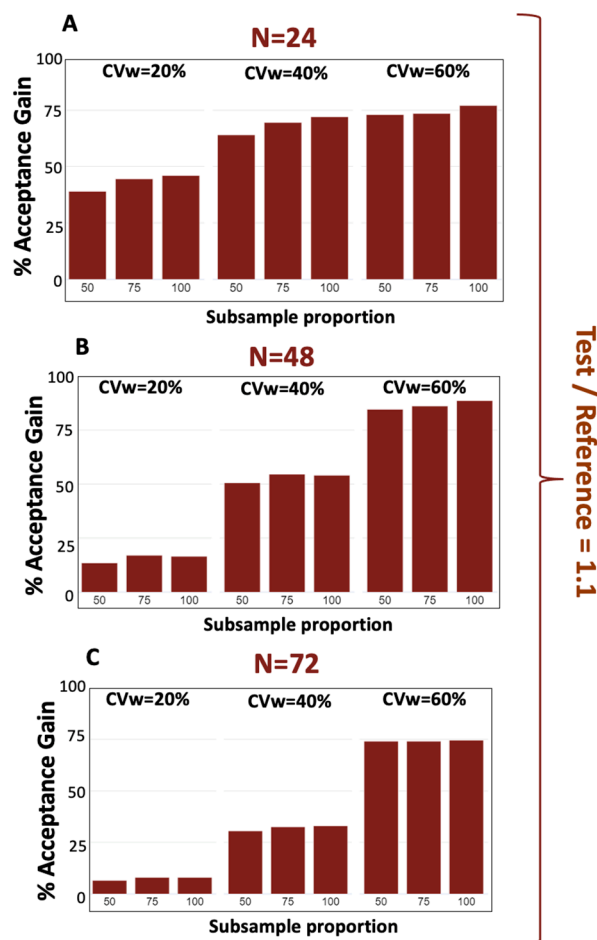


Figure 8. Acceptance gains between the VAE-generated and the original datasets when no scaling is applied to either dataset. Three different original sample sizes are used (24 (A), 48 (B), and 72 (C)), while the size of the generated datasets is equal (1×) to that of the original sample. In all cases, the average test/reference ratio is 1:1 and the within-subject variability equals 20%, 40%, and 60%.

From Figure 7, it becomes evident that in most scenarios there is large acceptance gain for the VAE method, which can reach high levels of up to 75% for $CV_W = 60\%$ and a sample size of 24. The lowest acceptance gain is observed when the CV is 40% and the sample size is 72 (i.e., 5%). There are also cases where there is no acceptance gain—that is, when the CV equals 20% for sample sizes of 48 and 72, and that is why these plots are omitted. Overall, the acceptance gain is much higher when the CV_W increases and the sample size decreases.

Figure 8 is similar to Figure 7, with the difference that the T/R now equals 1:1. Overall, the acceptance gain is higher as the CV increases and the sample size decreases, reaching a plateau at 75–80% when the CV equals 60%. Starting from 5% for a sample size of 72 (Figure 8C) and a 20% CV_W , the acceptance rate goes up to 80% for a sample size equal to 48 and a CV of 70% (Figure 8B). The acceptance gain is high overall (i.e., more than 30%) in the case where the sample size is 24 (Figure 8A).

4. Discussion

The aim of this study is to expand the application of a generative AI algorithm (in particular, VAEs) in the BE testing of highly variable drugs. High variability (“noise”) is a critical issue in the field of bioequivalence [1–3]. Highly variable drugs are medications that demonstrate considerable variability in their pharmacokinetic parameters, such as absorption, distribution, metabolism, and elimination processes, upon administration to individuals. Achieving strict bioequivalence for these drugs can be challenging due to their inherent pharmacokinetic variability. When conducting BE studies for such drugs,

the variability in individual responses can result in broader confidence intervals, making it more challenging to demonstrate equivalence within standard regulatory requirements. Thus, regulatory authorities often establish specific guidelines and acceptance criteria for highly variable drugs to accommodate this expected variability [1,2]. These criteria may involve widening acceptable ranges for certain pharmacokinetic parameters (e.g., C_{max}), utilizing alternative statistical methods for assessing bioequivalence (e.g., scaled equivalence), or increasing the sample size in studies [3]. Therefore, it is crucial to develop methods that reduce this undesired variability to avoid the necessity of increasing study participants, costs, and study complexity. The rationale for investigating VAEs in highly variable drugs lies in their inherent property of reducing variability.

To demonstrate the utility of VAEs in reducing the need for large sample sizes, Monte Carlo BE studies were simulated (5000 trials under each scenario) with and without the use of VAEs, and the statistical power in each case was recorded [3]. The performance of VAEs in the BE of highly variable drugs was compared against the classic 80–125% acceptance range and the scaled BE limits proposed by the US FDA and EMA [1,2]. Various scenarios were explored, focusing on high-variability values of the drugs under comparison. Also, several sample sizes, T/R ratios, and proportions of actual data used for generating the virtual data were further investigated. The performance of the VAE-generated dataset was compared with the original dataset and the subsampled dataset. We performed extensive hyperparameter tuning to enhance the efficiency of VAEs. In this context, we tested different combinations of parameters to assess the algorithm thoroughly. Preventing the generation of fake data is crucial, and in our research, we implemented all necessary steps to guarantee complete reproducibility of the entire process.

In all scenarios investigated in this study, the generated dataset exhibited significantly higher statistical power compared to the subsampled datasets, even when scaled acceptance limits were applied to the subsampled sets. It is noteworthy that the VAE-generated datasets consistently outperformed the much larger original datasets, sometimes by up to twice the statistical power. This trend persisted even when scaled limits were applied to the original dataset. Overall, the utilization of the VAE method resulted in performance that was at least equivalent to the much larger original dataset, and in many cases, it was notably superior. When no scaling was applied to the original dataset, i.e., when the VAE approach was treated equally to the other datasets, its superiority was even more pronounced.

More specifically, under the assumption that the ratio of the mean endpoints of the T and R groups was equal to 1 (indicating similar average performance between the two groups) with moderate within-subject variability (20%), the acceptance rate of the VAE-generated dataset was consistently above 95%, whereas the acceptance rate for the original dataset dropped to 80% for low sample sizes and subsampled proportions (Figures 3 and 5). With 40% within-variability and T/R equal to 1, the VAE-generated dataset exhibited performance comparable to that of the original dataset, despite the fact that scaling was applied to the original dataset, but unscaled limits were used in the VAE-synthesized datasets. Additionally, the generated dataset outperformed the original in the case of $CV_W = 60\%$. Under the assumption that T/R equals 1:1, and CV_W is 20% and 40%, the generated dataset performed at least as well as the original for high sample and subsample sizes but considerably better for low sample sizes. When within-variability was 60%, the VAE-generated dataset performed as well as the original when sample sizes were high but significantly better for the EMA case. In all cases, the generated dataset performed significantly better than the subsampled dataset, highlighting the added value of the VAE model. As expected, the original dataset showed much better performance than the subsampled dataset in all scenarios studied. Overall, it is important to note that in all scenarios presented above, no scaling was applied to the new VAE approach, but either the EMA or FDA scaled BE limits were applied to the original or subsampled data. Thus, if scaling were further applied to the VAE-generated dataset, the performance of the VAE would be incomparable to that of the original dataset and of course against that of the subsampled dataset.

Overall, these findings suggest that utilizing VAEs to virtually increase the sample size in BE studies for highly variable drugs can effectively enhance the statistical power of the study while requiring fewer subjects, thereby reducing human exposure, study completion time, costs, and overall study complexity. Also, the statistical analysis will be simpler and unified in all cases, since no scaling of BE limits or its a priori statement in the study protocol would be necessary. Additionally, it is worth noting that this study simulated the typical conditions of 2×2 crossover BE studies by integrating Monte Carlo simulations, stochastic terms, and VAEs. However, in practical application, simulations would not be necessary; VAEs could be applied directly to limited actual human data. The Monte Carlo simulations in this study were conducted solely to compare the performance of the VAE approach against the typical sample data used today.

A limitation of this study was the high computational time required for training the VAE model and the extensive number of scenarios tested, which constrained the study to 5000 runs. Another limitation is the use of the 2×2 typical crossover design. Thus, when the scaled BE approach was used for the “original” and “subsamped” datasets, the within-subject variability of the R drug was used in the reference scaled approach [1,2], namely, the stochastic term σ_w . For simplicity reasons, the same stochastic term σ_w was set for both T and R products, supposing that there is no reason to have different variabilities, i.e., assuming similar product performances.

It should also be mentioned that while Monte Carlo simulations are invaluable for testing hypotheses and understanding theoretical outcomes, the use of actual (real) data is crucial for validating these findings in real-world scenarios. Simulated data, although controlled and replicable, may not capture the complexities and nuances present in real-world situations. Actual data provide a more accurate and reliable basis for analysis, ensuring that the proposed methods are applicable and effective in practical applications. Moreover, using real data can reveal unforeseen challenges and factors that simulations might overlook, leading to more robust and generalizable conclusions. Therefore, incorporating real data is essential to move from theoretical exploration to practical implementation and to ensure the relevance and impact of the research findings.

In this work, VAEs are proposed as a method to augment data in clinical trials, specifically to potentially reduce the necessary sample size. VAEs extend traditional autoencoders by utilizing an encoder network that maps input data to a multivariate normal distribution instead of a fixed point. Essentially, VAEs aim to establish a relationship between input data and a probability distribution across the latent space [26]. The adoption of VAEs brings numerous benefits. They have been recognized as a highly effective approach for developing generative models, which generate new synthetic or artificial data based on existing data. A significant advantage of VAEs over conventional autoencoders is their capability to generate new data from the same underlying distribution as the input data by sampling from this embedded distribution. These distinctive characteristics were used in this study to effectively reduce the actual sample size of a BE study.

Significantly more work needs to be carried out on the integration of AI algorithms in the BE assessment of highly variable drugs. This study is only the first step towards this goal. Future studies should explore additional clinical designs commonly used for highly variable drugs, such as the semi-replicate 3×3 or replicate designs like 2×4 . It is important to highlight that VAE-synthesized data cannot be utilized for safety assessments, as these require actual human data, but only as a means for increasing the statistical power of the study. The use of generative AI algorithms in clinical research necessitates rigorous validation and comprehensive ethical review of the synthetic data they produce [18]. The validation process should encompass comparison with real-world data, performance benchmarks, and external validation. Initially, extensive studies should compare synthetic data characteristics to actual clinical trial data, utilizing statistical analyses to ensure fidelity and representativeness. Subsequently, the generative model should be validated against established datasets using industry-standard metrics to confirm accuracy and reliability. Collaboration with independent researchers for external validation is essential to ensure

objectivity and transparency. Detailed documentation of the data generation and validation process should be available for review by ethical committees. Ethical considerations also include obtaining consent for using clinical data in generating synthetic datasets and addressing potential biases to ensure fairness and representativeness. Adhering to these steps will ensure that the synthetic data generated by generative AI algorithms are scientifically valid and ethically sound for clinical research [21].

As AI continues to advance and its applications in healthcare proliferate, its value as a tool becomes increasingly evident [21,30–34]. In our approach of using AI to virtually increase the sample size of BE studies, the benefits in terms of enhanced statistical power while using fewer human subjects become evident. However, it is important to emphasize that the proposed methodology does not advocate for the complete replacement of subjects in BE studies but rather supplements them. It is crucial to have an appropriate and representative sample size of actual human subjects to generate synthetic ones. This study demonstrates that VAEs can reduce the originally required sample size by up to 50% while maintaining or even increasing statistical power. This new idea suggests synthesizing new virtual subjects from a limited number of existing ones, therefore increasing statistical power without increasing human exposure and resources and avoiding ethical concerns, costs, and time constraints.

Another important aspect of this new approach is its full reproducibility [34–37]. By appropriately training the VAE neural network and using seeds for all random generations, the synthetic data can increase the sample size in a cost-free manner without the limitations of other approaches like bootstrapping. While techniques like bootstrapping have historically been considered for augmenting sample sizes, they are ultimately limited by the repetition of subjects, which can introduce bias and overfitting. Therefore, new methods like the one proposed here are essential. Finally, it is crucial for regulatory authorities such as the EMA and the FDA to establish criteria and guidelines for the use of AI-generated virtual subjects in studies, similar to the regulations set for pharmacometrics models around 10 years ago [38–40]. This will ensure the reliability and validity of studies utilizing synthetic data and promote the responsible integration of AI into healthcare research.

5. Conclusions

This study focuses on the use of VAEs in BE trials of highly variable drugs. In this context, we simulated various scenarios, including different high variability levels, by appropriately setting the stochastic terms of the model, the actual sample sizes of the BE study, AI-generated sample sizes, and the relationship between the average performance of the T and R products. The use of artificial neural networks, particularly VAEs, to reduce the need for recruiting large numbers of subjects in BE studies of highly variable drugs offers numerous advantages, including significantly reduced human exposure, shorter study completion times, simplified trial processes, less workload for healthcare professionals, considerably lower costs for sponsors, and less complexity in the statistical analysis since no scaled BE limits will be necessary. Overall, this study advocates for the integration of AI-driven generative algorithms in clinical research. Implementing these innovative concepts in practice would require regulatory authorities to establish specific criteria and guidelines to ensure the proper application of AI-generated virtual subjects, thereby avoiding potential issues such as hallucinations and ensuring reproducibility.

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

Funding: G.K. is supported in the framework of H.F.R.I. called “Basic Research Financing (Horizontal Support of All Sciences)” under the National Recovery and Resilience Plan “Greece 2.0” funded by the European Union—NextGenerationEU. (H.F.R.I. project number: 14910).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

Appendix A

Neural networks, autoencoders, and variational autoencoders

Neural networks are machine learning models, which have gained popularity in recent years due to the accessibility of cloud computing and their vast range of applications. They are used for solving both supervised and unsupervised problems, classification, regression, clustering, and data augmentation; therefore, different neural network architectures exist, each of which is optimally suited for addressing these problems. The cornerstone of neural networks is the neuron. A vertical arrangement of several neurons forms a layer and a horizontal arrangement of many layers forms the neural network. The first layer of the neural network is referred to as the “input” layer, whereas the last layer is called the “output” layer. The layers in between the input and output layers are referred to as “hidden”.

Autoencoders (AEs) are neural networks that are usually used for input reconstruction, which is possible due to their unique architecture. They are composed of two distinct elements, the encoder and the decoder, which are linked via a latent space. Both the encoder and the decoder consist of the same number of hidden layers, but their order is reversed; thus, the decoder is a mirrored image of the encoder. The input data flow in a left-to-right manner from the encoder and are mapped into a fixed-point representation in a latent space. A dimension of the latent space that is smaller than the input’s dimension will result in an “incomplete” autoencoder, whereas a larger dimension of the latent space compared to the input’s dimension will result in an “overcomplete” autoencoder. The decoder receives as input the output of the encoder and forward propagation is performed, the output is compared to the input and the difference (i.e., reconstruction loss) is calculated, and throughout the backpropagation the error is minimized.

VAEs are an extension of conventional autoencoders. Instead of a fixed-point representation, the latent space is a probability distribution (typically Gaussian), where its mean and variance are calculated throughout the training of the neural network. This is achieved by adding the Kullback–Leibler loss to the reconstruction loss of AEs. VAEs are “generative” algorithms due to the ability to randomly sample data from the latent space, pass the sampled data through the decoder, and generate novel datapoints that are “similar” to the input data. The general architecture of a variational autoencoder is presented in Figure A1.

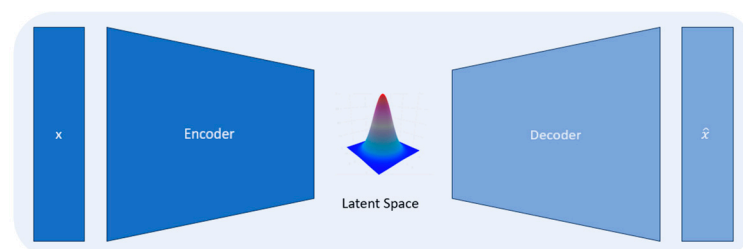


Figure A1. Visual representation of the architecture of a variational autoencoder. The input data x flow from left to right from the encoder and are mapped to the latent space. The decoder processes the output of the encoder and produces the output $\hat{x}$.

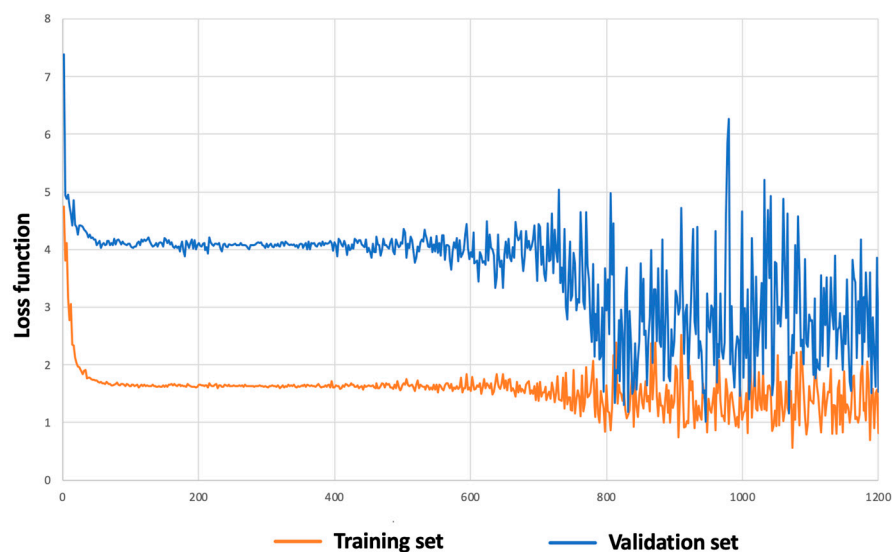


Figure A2. Loss function as a function of the number of epochs for the training and validation sets.

Number of epochs

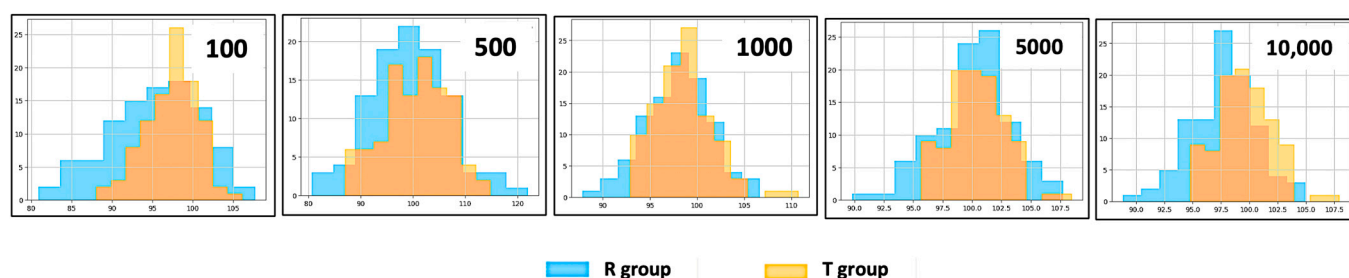


Figure A3. The generated data distribution for both the reference (R) and test (T) groups synthesized from variational autoencoders. The five graphs refer to the T and R distributions across different numbers of epochs: 100, 500, 1000, 5000, and 10,000. The orange color represents the data points that correspond to both T and R distributions.

References

- EMA. Rev. 1/Corr **: Committee for Medicinal Products for Human Use (CHMP). Guideline on the Investigation of Bioequivalence. 2010. Available online: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf (accessed on 14 April 2024).
- Guidance for Industry: Bioavailability and Bioequivalence Studies Submitted in NDAs or INDs—General Considerations. Draft Guidance. 2014. Available online: <https://www.fda.gov/media/88254/download> (accessed on 14 April 2024).
- Karalis, V. Modeling and Simulation in Bioequivalence. In *Modeling in Biopharmaceutics, Pharmacokinetics and Pharmacodynamics. Homogeneous and Heterogeneous Approaches*, 2nd ed.; Springer International Publishing: Cham, Switzerland, 2016; pp. 227–255.
- Noordzij, M.; Dekker, F.W.; Zoccali, C.; Jager, K.J. Sample Size Calculations. *Nephron Clin. Pract.* **2011**, *118*, c319–c323. [\[CrossRef\]](#)
- Flight, L.; Julious, S.A. Practical Guide to Sample Size Calculations: Non-Inferiority and Equivalence Trials. *Pharm. Stat.* **2015**, *15*, 80–89. [\[CrossRef\]](#) [\[PubMed\]](#)
- Shih, W.J.; Li, G.; Wang, Y. Methods for Flexible Sample-Size Design in Clinical Trials: Likelihood, Weighted, Dual Test, and Promising Zone Approaches. *Contemp. Clin. Trials* **2016**, *47*, 40–48. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zheng, C.; Wang, J.; Zhao, L. Testing Bioequivalence for Multiple Formulations with Power and Sample Size Calculations. *Pharm. Stat.* **2012**, *11*, 334–341. [\[CrossRef\]](#)
- Tubert-Bitter, P.; Manfredi, R.; Lellouch, J.; Bégaud, B. Sample Size Calculations for Risk Equivalence Testing in Pharmacoepidemiology. *J. Clin. Epidemiol.* **2000**, *53*, 1268–1274. [\[CrossRef\]](#)

9. D'Arrigo, G.; Roumeliotis, S.; Torino, C.; Tripepi, G. Sample Size Calculation of Clinical Trials in Geriatric Medicine. *Aging Clin. Exp. Res.* **2020**, *33*, 1209–1212. [[CrossRef](#)] [[PubMed](#)]
10. Tang, B.-H.; Yao, B.-F.; van den Anker, J.; Zhao, W. Optimal Sample Size for Use in Neonatal Pharmacokinetic Studies. *Ther. Innov. Regul. Sci.* **2022**, *56*, 517–522. [[CrossRef](#)] [[PubMed](#)]
11. 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](#)]
12. Martin-McGill, K.J.; Bresnahan, R.; Levy, R.G.; Cooper, P.N. Ketogenic Diets for Drug-Resistant Epilepsy. *Cochrane Database Syst. Rev.* **2020**, *6*, CD001903. [[CrossRef](#)] [[PubMed](#)]
13. Hajian-Tilaki, K. Sample Size Estimation in Diagnostic Test Studies of Biomedical Informatics. *J. Biomed. Inform.* **2014**, *48*, 193–204. [[CrossRef](#)]
14. 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](#)] [[PubMed](#)]
15. 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](#)]
16. 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](#)]
17. Ramesh, A.; Kambhampati, C.; Monson, J.; Drew, P. Artificial Intelligence in Medicine. *Ann. R. Coll. Surg. Engl.* **2004**, *86*, 334–338. [[CrossRef](#)] [[PubMed](#)]
18. Ossowska, A.; Kusiak, A.; Świetlik, D. Artificial Intelligence in Dentistry—Narrative Review. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3449. [[CrossRef](#)] [[PubMed](#)]
19. Hashimoto, D.A.; Witkowski, E.; Gao, L.; Meireles, O.; Rosman, G. Artificial Intelligence in Anesthesiology. *Anesthesiology* **2019**, *132*, 379–394. [[CrossRef](#)] [[PubMed](#)]
20. Keskinbora, K.; Güven, F. Artificial Intelligence and Ophthalmology. *Turk. J. Ophthalmol.* **2020**, *50*, 37–43. [[CrossRef](#)] [[PubMed](#)]
21. Karalis, V.D. The Integration of Artificial Intelligence into Clinical Practice. *Appl. Biosci.* **2024**, *3*, 14–44. [[CrossRef](#)]
22. Papadopoulos, D.; Karalis, V.D. Variational Autoencoders for Data Augmentation in Clinical Studies. *Appl. Sci.* **2023**, *13*, 8793. [[CrossRef](#)]
23. Abdi, H.; Valentin, D.; Edelman, B. *Neural Networks*; SAGE University Papers Series: Quantitative Applications in the Social Sciences; Sage: Newcastle upon Tyne, UK, 1999.
24. Bishop, C.M. Neural Networks and Their Applications. *Rev. Sci. Instrum.* **1994**, *65*, 1803–1832. [[CrossRef](#)]
25. Müller, B.; Reinhardt, J.; Strickland, M.T. *Neural Networks: An Introduction*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
26. Doersch, C. Tutorial on Variational Autoencoders. *arXiv* **2016**, arXiv:1606.05908. Available online: <https://arxiv.org/pdf/1606.05908.pdf> (accessed on 14 April 2024).
27. 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](#)]
28. Kingma, D.P.; Welling, M. An Introduction to Variational Autoencoders. *Found. Trends® Mach. Learn.* **2019**, *12*, 307–392. [[CrossRef](#)]
29. Yu, T.; Zhu, H. Hyper-Parameter Optimization: A Review of Algorithms and Applications. *arXiv* **2020**, arXiv:2003.05689. Available online: <https://arxiv.org/abs/2003.05689> (accessed on 14 April 2024).
30. Polevikov, S. Advancing AI in Healthcare: A Comprehensive Review of Best Practices. *Clin. Chim. Acta* **2023**, *548*, 117519. [[CrossRef](#)]
31. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial Intelligence in Healthcare: Past, Present and Future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [[CrossRef](#)] [[PubMed](#)]
32. Koski, E.; Murphy, J. AI in Healthcare. *Stud. Health Technol. Inform.* **2021**, *284*, 295–299. [[CrossRef](#)]
33. Chen, R.J.; Lu, M.Y.; Chen, T.Y.; Williamson, D.F.K.; Mahmood, F. Synthetic Data in Machine Learning for Medicine and Healthcare. *Nat. Biomed. Eng.* **2021**, *5*, 493–497. [[CrossRef](#)]
34. Mahmoud, A.Y.; Neagu, D.; Scrimieri, D.; Abdullatif, A.R.A. Early Diagnosis and Personalised Treatment Focusing on Synthetic Data Modelling: Novel Visual Learning Approach in Healthcare. *Comput. Biol. Med.* **2023**, *164*, 107295. [[CrossRef](#)]
35. Foster, D. *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*, 2nd ed.; Karl Friston (Foreword) Oreilly & Associates Inc.: London, UK, 2023.
36. 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](#)]
37. Chien, J.-T. Deep Neural Network. In *Source Separation and Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 259–320.
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: Silver Spring, MD, USA; Center for Drug Evaluation and Research (CDER): Silver Spring, MD, USA; Center for Biologics Evaluation and Research (CBER): Las Vegas, NV, USA, 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): Amsterdam, The Netherlands, 2007. Available online: <https://www.ema.europa.eu/en/reporting-results-population-pharmacokinetic-analyses-scientific-guideline> (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.