


Variational Autoencoders for Data Augmentation in Clinical Studies

Dimitris Papadopoulos¹ and Vangelis D. Karalis^{1,2,*} 

¹ Department of Pharmacy, School of Health Sciences, National and Kapodistrian University of Athens, 15784 Athens, Greece

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: Variational autoencoders, which are a type of neural network, are introduced in this study as a means to virtually increase the sample size of clinical studies and reduce costs, time, dropouts, and ethical concerns. The efficiency of variational autoencoders in data augmentation is proven through simulations of several scenarios.

Abstract: Sample size estimation is critical in clinical trials. A sample of adequate size can provide insights into a given population, but the collection of substantial amounts of data is costly and time-intensive. The aim of this study was to introduce a novel data augmentation approach in the field of clinical trials by employing variational autoencoders (VAEs). Several forms of VAEs were developed and used for the generation of virtual subjects. Various types of VAEs were explored and employed in the production of virtual individuals, and several different scenarios were investigated. The VAE-generated data exhibited similar performance to the original data, even in cases where a small proportion of them (e.g., 30–40%) was used for the reconstruction of the generated data. Additionally, the generated data showed even higher statistical power than the original data in cases of high variability. This represents an additional advantage for the use of VAEs in situations of high variability, as they can act as noise reduction. The application of VAEs in clinical trials can be a useful tool for decreasing the required sample size and, consequently, reducing the costs and time involved. Furthermore, it aligns with ethical concerns surrounding human participation in trials.



Citation: Papadopoulos, D.; Karalis, V.D. Variational Autoencoders for Data Augmentation in Clinical Studies. *Appl. Sci.* **2023**, *13*, 8793. <https://doi.org/10.3390/app13158793>

Academic Editors: Qingchen Zhang and Jinyang Huang

Received: 2 June 2023

Revised: 25 July 2023

Accepted: 28 July 2023

Published: 30 July 2023



Copyright: © 2023 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/>).

Keywords: variational autoencoders; clinical trials; data augmentation; sample size

1. Introduction

Sample size estimation is a crucial component of clinical trials since the latter serves as the cornerstone for ensuring safety and efficacy [1]. A representative sample of an adequate size can provide insights into a given population. However, the collection of substantial amounts of data may prove challenging, costly, and time-intensive. It is imperative that each clinical trial be carefully organized through the development of a protocol that outlines the study's objectives, primary and secondary endpoints, data collection methodology, sample selection criteria, data handling procedures, statistical methods and assumptions, and, on top of that, a scientifically justified sample size [1].

The determination of sample size can vary significantly based on the study design, outcome type, and hypothesis test specified by the investigator [2]. The estimation of appropriate sample size is based on the given statistical hypotheses and several study design parameters. The aforementioned factors encompass the minimal detectable difference that holds meaning, estimated variability in measurement, desired level of statistical power, and level of significance [2]. Achieving an optimal balance between an insufficient or excessive number of participants in the sample is imperative [3]. Insufficient statistical power resulting from a small sample size may lead to a failure to detect a true difference, thereby

rendering significant variations among study groups statistically insignificant. The utilization of an excessively large sample size can be deemed unethical, result in the wasteful use of resources, and potentially impede the feasibility of a given study. Furthermore, there is a growing expectation from funding agencies, ethics committees, and scientific journals for the justification of sample size. In certain scenarios, such as the evaluation of highly variable drugs in bioequivalence assessment, it is imperative to utilize large sample sizes as specified by regulatory bodies such as the EMA in 2010 and the FDA [4–6]. Regardless of the underlying cause, when variability increases, demonstrating bioequivalence becomes more challenging, despite its existence. In general, as the degree of variability increases, it becomes increasingly challenging to prove what is sought unless a larger sample size is employed.

In this context, computational alternatives can be found for increasing the sample size and thus decreasing human exposure [4]. Data augmentation is a methodology employed to expand the sample by generating modified replicas of a given dataset through the utilization of pre-existing data. This involves implementing slight modifications to the dataset or utilizing deep learning techniques to produce novel data instances. The potential of artificial intelligence (AI) as a means to achieve sustainable and enhanced drug development has been acknowledged, leading to the exploration and discussion of various applications in clinical trials [7]. The significance of data availability in the context of data-driven and individualized healthcare trends cannot be overstated. However, the process of producing useful insights from accessible data necessitates the utilization of comprehensive AI models. These models must be built and trained using suitable datasets to effectively accelerate and simplify every step within drug research, as has been noted in previous studies [8,9].

Recently, the Alan Turing Institute initiated a project whose primary objective was to investigate the potential influence of machine learning and artificial intelligence on the planning, implementation, and interpretation of randomized clinical trials [10]. Through the augmentation of human expertise and optimization of data utilization, artificial intelligence has the capability to forecast the probability of trial or site failure, as well as clinical patient outcomes. AI can also be used to analyze health records to identify appropriate cohorts for clinical trials, accelerate trial recruitment, monitor clinical trials, and effectively notify medical personnel and patients about available trial opportunities [10]. Additionally, the simplification of entry criteria can enhance accessibility for potential participants.

In this vein, autoencoders have been identified as a highly effective and valuable technique for generating synthetic or artificial data from real-world data. The autoencoder is a variant of the artificial neural network that is employed for acquiring proficient codifications of unannotated data [11]. The autoencoder acquires knowledge through two distinct functions: an encoding function that modifies the input data and a decoding function that reconstructs the input data from the encoded representation. Variational autoencoders (VAEs) are a type of generative model that operates under a probabilistic framework and incorporates neural networks as a component of its broader architecture [12,13].

The aim of this study is to introduce a new data augmentation idea in clinical trials by using VAEs to reduce the required sample size. In order to achieve this task, several forms of VAEs were explored and used for the generation of virtual populations. The VAE-generated subjects were appropriately set up in the form of an equivalence study. The first step of this analysis was tuning the VAE system by selecting the most appropriate hyperparameters. In the next step, the previously tuned VAE model was used to explore several scenarios between the assessments of two groups of volunteers (i.e., test (T) vs. reference (R)).

2. Materials and Methods

Artificial intelligence refers to the replication of human intelligence in machines and computer systems. AI involves the development of intelligent machines that possess the ability to perform tasks that are either equivalent to or surpass those of human beings [14].

The procedure involves gathering data, developing usage guidelines, arriving at approximations or conclusions, and self-correcting (i.e., minimizing errors during training). In several fields, including pharmaceutical sciences, both AI and machine learning approaches have drawn a lot of interest. Deep learning involves learning successive layers of representations that are more and more pertinent to the data [15]. The models known as neural networks, which are organized in layers piled on top of one another, are used in deep learning to learn these layered representations (nearly always). The most cutting-edge deep learning techniques take advantage of recent advances in neural networks. Compared to other machine learning techniques, they frequently have better prediction and generalization skills. The discovery of drugs and repurposing are two fields of pharmaceutical research where deep learning has attracted interest and gained recognition [16].

2.1. Strategy of the Analysis

Classically, in clinical trials, the sample size is explicitly stated in the study protocol and, therefore, estimated before the initiation of the study. Estimation of the sample size is a complex procedure that relies on several parameters of the trials, among which the most important are measured endpoint (or endpoints), measurement scale of the endpoint, variability of the endpoint, nominal level of the type I error, maximum anticipated type II error, acceptance limits, and difference between the treatments (in the case of interventional studies). However, when the variability of the endpoint(s) is high, the type I error is set to be very low or we want to increase the statistical power of the study (i.e., decrease type II error), there is a need for a large sample size. The latter leads to the inclusion of many human participants, rather increased costs, a long duration of the study, a high possibility of dropouts, etc. There is not much we can do to limit the sample size; only in the case of bioequivalence studies, the scaled average approach has been proposed, where the acceptance limits scale as a function of the residual variability of the study [4–6].

The aim of this work is to introduce a novel idea for reducing the required sample size in clinical trials. The idea is as follows:

- (a) Perform the clinical study using a limited number of volunteers;
- (b) Using the results from “a”, apply in the next step a VAE in order to create virtual subjects and increase the statistical power.

The ideal situation would be one where we could achieve high statistical power without increasing the false positive rate (type I error). In the following lines, the methodology and results are presented of such a method where the latter requirements are fulfilled. In order to show that VAE works efficiently, an experimental method was set up. In brief, the experimental part is outlined below:

- i. Create N virtual subjects (e.g., $N = 100$) using Monte Carlo simulations. This is considered the “original” dataset.
- ii. Set the average endpoint value equal to 100 units and conduct sampling assuming log-normal distribution [4]. Several levels of variability (e.g., 10%, 20%, 40%, etc.) are used for the random creation of virtual subjects.
- iii. Assume two treatments: Test (T) and Reference (R), as in the case of bioequivalence studies. Several levels of the T/R ratios are explored.
- iv. Use these virtual subjects to form a clinical trial; for the purposes of this work, a parallel clinical design was used. Half of the subjects are considered to receive one treatment (e.g., T) and the other half received the other (e.g., R).
- v. Draw a random sample from the original dataset (steps “i” and “ii”) to create the sub-sample.
- vi. Apply VAE to the sub-sample created in the previous step (i.e., “v”). This leads to the creation of the “generated sample of subjects”.
- vii. Apply the typical statistics imposed by the regulatory authorities [5,6]. The statistical analysis compares T vs. R separately for the “original dataset”, “sub-sample”, and “VAE-generated dataset” (or simply the “generated dataset”).

- viii. Record the success or failure of the study separately for the “original dataset”, “sub-sample”, and “generated dataset”.
- ix. Repeat steps “i”–“viii” many times (e.g., 500) to obtain robust estimates for the percentage of acceptance (i.e., % success) of each of the three datasets.
- x. Compare the performances obtained in step “ix”.

The set-up, validation, and fine-tuning of the VAE hyperparameters were conducted exhaustively after step “vi”.

Generally speaking, it would be preferred for the VAE-generated dataset to result in higher percentages compared to the sub-sampled dataset. It would be almost ideal if the performance of VAE-generated data were similar to that of the original dataset. Finally, it would be ideal if the performance of the VAE dataset were even better than the original data. As will be shown later in this work, the latter exists, namely, the performance of the VAE-generated data was even better than when using the original data in cases of high variability.

2.2. Neural Networks—Autoencoders

In recent years, neural networks have gained popularity as state-of-the-art machine learning models. There are multiple neural network architectures, each of which is optimally suited for addressing distinct problem types. The fully connected neural network is widely regarded as the most prevalent type of neural network.

The fundamental component of a neural network is the neuron, which is also referred to as a node [11]. The vertical arrangement of numerous neurons creates a layer. In the context of neural networks, the initial stratum is commonly referred to as the input layer, while the final stratum is designated as the output layer. The number of nodes in the input layer corresponds to the number of features in the model, while the number of nodes in the output layer corresponds to the desired output. The intermediary strata within neural network architecture are commonly referred to as hidden layers. The number of hidden layers may vary based on the intricacy of the problem at hand. The variability of the number of neurons in each hidden layer is contingent upon the intricacy of the problem at hand. The determination of the optimal number of hidden layers and neurons is not governed by a universal principle. Typically, the determination of such a choice is derived from practical knowledge and trial and error, and it tends to vary based on the specific dataset and problem at hand.

The determination of parameters and the number of layers in a neural network is a crucial aspect of designing an effective and efficient model. In this study, these architectural choices were made based on the following: (a) starting with and focusing on simple architectures (for example, for both the encoder and the decoder, we tried to have a low number of hidden layers and number of neurons per layer), (b) performing hyperparameter tuning to find the optimal values for hyperparameters (see Section 2.4), and (c) using regularization techniques (like dropout and weight decay) to avoid overfitting.

The process of training a neural network involves sequential observations passing through the network, followed by forward and backward propagation [11]. The process entails sequentially transmitting the input, from left to right, through the network’s layers. Each layer executes a basic calculation (transformation) on the input and transmits the resulting output to the subsequent layer. The computation involves a linear combination of the input for every layer, which is weighted. Subsequently, an activation function, also referred to as non-linear transformation, is applied prior to passing the output to the next layer. The aforementioned procedure is iteratively executed until the input data traverse the entirety of the neural network. Upon completion of the forward propagation process, the error is computed by measuring the discrepancy between the output and the anticipated output in relation to the loss function.

The method of modifying the biases and weights of a network is known as backward propagation [11]. The ultimate objective of this particular stage is to reduce the cost function value. The definition of the cost value in our application was the typical loss function, where

the Kullback-Leibler (KL) loss part and the reconstruction part were equally weighted. The gradient of the error function with respect to the weights was used to update the weights during the training process. This was done in a right-to-left manner, where the end goal was to minimize the error function by adjusting the weights in small incremental steps. The complete iteration of both forward and backward propagation over the entire dataset is referred to as an “epoch” in the context of machine learning. Upon the completion of every epoch, the loss function, which represents the error, is computed. The objective is to select a suitable quantity of epochs such that the error attains convergence. The number of epochs may fluctuate based on the intricacy of the problem at hand.

Autoencoders (AEs) are a distinct class of neural networks that are commonly trained to perform input reconstruction [12,13]. Various architectures of autoencoders exist and are comprised of two distinct components, namely, the encoder and the decoder. The process of encoding involves the mapping of the input information into a fixed-point representation in a latent space. If the dimensionality of the latent space is lower than that of the input data, the encoder will acquire a more parsimonious representation of the input data, resulting in an incomplete autoencoder, namely, the number of neurons in the bottleneck layer will be smaller than the number of neurons in the input and output layers. On the other hand, if the latent dimension exceeds the input dimension, the autoencoder acquires a superfluous depiction of the input data, resulting in an overcomplete autoencoder, namely, the number of neurons in the bottleneck layer will be larger than the number of neurons in the input and output layers.

The utilization of autoencoders can serve two distinct purposes. Firstly, in one scenario, autoencoders can be employed to reduce noise in a given dataset. Secondly, in another scenario, autoencoders can be utilized to identify intricate patterns within the input data, which can subsequently enhance the accuracy of their reconstruction. The decoder component receives the encoder output as its input and performs forward propagation until it gets to the output layer. At this stage, the error is computed in relation to the chosen loss function. The decoder possesses identical dimensions to those of the encoder, albeit in a mirrored orientation [12,13]. Autoencoders undergo training via forward and backward propagation, similar to other neural networks. Throughout the training process, the autoencoder endeavors to minimize the error in reconstruction that exists between the input and the rebuilt output. In neural networks, loss functions are used to quantify the difference between the predicted outputs of the neural network and the actual target values. In this study, the convergence of the algorithm was tested with respect to the value of the loss function, which was computed at the end of each iteration (epoch). This means that the value of the loss function at the last epoch was stabilized.

2.3. Variational Autoencoders

Variational autoencoders represent an expansion of conventional AEs [12,13]. In conventional autoencoders, the encoder acquires a latent representation of the data, which is subsequently utilized by the decoder to reconstruct the initial input data. The process is executed deterministically, indicating that identical input will yield identical output. In contrast to other methods, VAEs aim to establish a mapping between the input data and a probability distribution across the latent space. Specifically, this distribution is represented by the mean and variance of a Gaussian distribution, which is typically utilized. The ability to perform random sampling from the latent space is a valuable technique as it enables the subsequent utilization of this output as an input for the decoder component, thereby facilitating the generation of new data.

The primary aim of VAEs is to reduce the reconstruction loss, which is similar to that of a conventional AE. However, VAEs also strive to minimize the KL differences between the acquired distribution and a prior distribution across the latent space. The KL divergence quantifies the degree of resemblance between two probability distributions, especially the extent to which distribution Q provides an adequate approximation of distribution P. Assuming that x refers to the input data and z represents the latent variables or the encoded

representation of the input data, the objective in the context of VAE is to approximate the posterior distribution $P(z|x)$, which facilitates the projection of data into the latent space. Due to the unknown nature of $P(z|x)$, a simplified estimation of $Q(z|x)$ is utilized. In the process of training a VAE, the encoder module is trained to minimize the discrepancy between the posterior distribution $Q(z|x)$ and the prior distribution $P(z|x)$ by optimizing the KL divergence between the two distributions. Consequently, the objective function of the VAE comprises the divergence of the KL term, necessitating its minimization.

2.4. Tuning of Hyperparameters

In a neural network, such as a VAE, the most important hyperparameters refer to the number of hidden layers, number of neurons per hidden layer, number of epochs, activation function, optimization function, weight initialization, dropout rate, and regularization. In this study, the optimization of hyperparameters was performed using a grid search. In particular, several sets of values were predefined, and the performance of the VAE model was exhaustively evaluated through an iterative process of trial and error [11]. Various values were tested for epochs, such as 100, 500, 1000, 5000, and 10,000. In relation to the activation function, the experiments were conducted utilizing both “softplus” and “linear” activation functions for both the hidden and output layers. The “softplus” activation function is a smooth, nonlinear activation function that converts the input (logit) into a positive range. The “softplus” function is defined by Equation (1):

$$\text{“Softplus” activation function} = \log(1 + e^x) \quad (1)$$

The “softplus” function is widely used since it is continuous and differentiable everywhere. Also, its smoothness helps in training neural networks more effectively through gradient-based optimization algorithms, such as gradient descent. Different methods were used to assess which hyperparameters better fit our application. This included the percentage of convergence, degree of similarity between the generated data, and original bell-shaped source data.

The KL component and the reconstruction component of the loss function were equally weighted, and the dimension of the latent space was chosen. Finally, with regard to the number of hidden layers and the number of neurons in each hidden layer, various configurations were explored consisting of 2, 3, and 4 hidden layers for both the encoder and decoder. The encoder consisted of 128, 64, 32, and 16 neurons, while the decoder consisted of 16, 32, 64, and 128 neurons, respectively. Table 1 displays all of the aforementioned factors tested during the development of the VAE.

Table 1. Hyperparameter tuning during the development of the variational autoencoders. In all cases, the latent space dimension was equal to 1.

Number of Epochs	Activation Function		Weights of Loss Function		Number of Hidden Layers		Number of Neurons in Hidden Layers (from Left to Right)	
	Hidden Layers	Output Layer	KL Part	Reconstruction Part	Encoder	Decoder	Encoder	Decoder
100	softplus	softplus	1	1	2	2	32-16	16-32
500	linear	linear			3	3	64-32-16	16-32-64
1000					4	4	128-64-32-16	16-32-64-128
5000								
10,000								

Key: KL, Kullback–Leibler difference.

All possible combinations of the factors listed on the left-hand side of Table 1 were investigated. The experimental runs detailed in Table 1 utilized the TensorFlow 2.10.0 Python

package and Python version 3.7, with execution taking place within a “Jupyter notebook” environment. Figure 1 depicts the general architecture of a variational autoencoder.

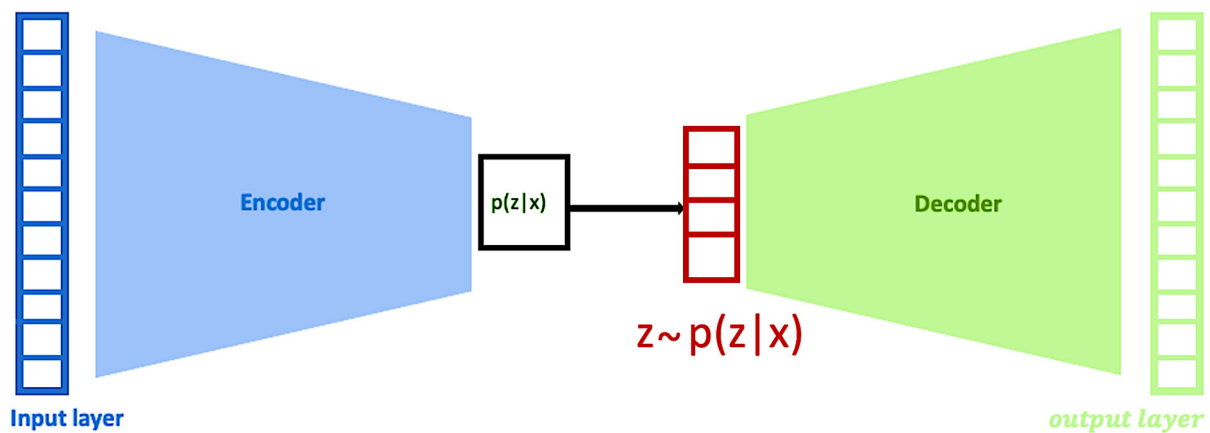


Figure 1. Visual representation of a variational autoencoder. The process of encoding involves compressing data from their original space to a latent space, while the decoding process involves decompressing the data. The methodology involves the utilization of neural networks as both an encoder and a decoder, with the aim of acquiring an optimal encoding–decoding scheme through an iterative optimization process. Variational autoencoders aim to establish mapping between the input data and a probability distribution across the latent space.

2.5. Monte Carlo Simulations

The methodology employed for the generation of subjects was as follows: Initially, a sample of 100 subjects was generated for the reference (i.e., R) group through a random process, utilizing a normal distribution with a mean of μ_R (i.e., the average endpoint value) and a standard deviation of σ_R . Then, a random subsampling procedure was performed on the original R group, whereby a proportion (gradually decreasing from 90% to 10% with a step of 10%) of the data, termed “subsample size”, was selected from the distribution. Subsequently, the subsample was utilized to train the VAE model, followed by sampling from the inferred latent distribution and generating a total of 100 virtual subjects for the R group. Similarly, the aforementioned procedure was also repeated for the test (i.e., T) group of subjects. In the case of the T group, random generation was based on a mean endpoint value of μ_T and standard deviations (σ_T). The aforementioned procedure is schematically shown in Figure 2.

Several ratios between the average endpoint values of the T and R groups were explored. In order to achieve this task, the mean endpoint value for R was set at 100, while for the T group, it was equal to $1.00\times$, $1.10\times$, $1.25\times$, and $1.50\times$ times that of R. In all cases, the coefficient of variation (CV) was equal between the T and R groups, and three different CV values were explored: 10% (low variability), 20% (medium variability), and 40% (high variability).

According to the aforementioned procedure, three different types of samples were utilized in the simulations: (a) the original dataset, (b) the subsampled group, and (c) the regenerated group by using the VAE system. To find out how well the VAE approach works, statistical analyses were conducted to compare the original, subsampled, and made-up data within and between the R and T groups. In addition, comparisons were made between the T and R groups of each dataset, namely, the T vs. R of the original dataset, the T vs. R of the sub-sampled dataset, and the T vs. R of the VAE dataset.

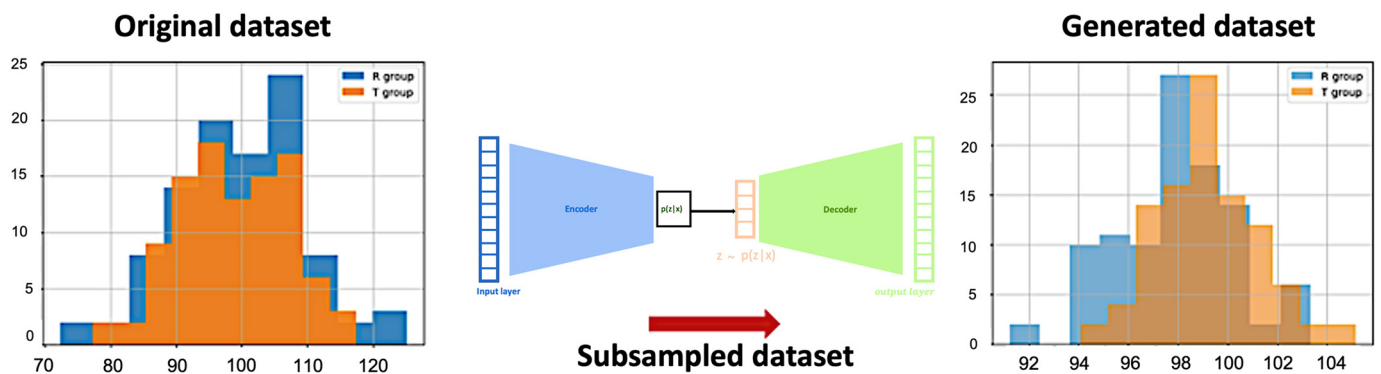


Figure 2. Schematic representation of the analysis strategy in this study. Initially, two randomly generated datasets were generated for the test (T) and reference (R) groups. Then followed subsampling to draw parts of the original population. Finally, the variational autoencoder was applied to the subsampled data in order to produce the generated datasets. The aim of the generated datasets was to exhibit the same properties as the original data. In this study, comparisons were made among the three datasets (original vs. subsampled vs. generated), as well as between the T and R groups of all datasets.

After the generation of the virtual subjects, they were appropriately classified into two groups in order to construct a parallel clinical design [4]. Ln-transformation (Napierian) was applied to the generated values before proceeding to the statistical analysis. Since the data obtained from bioequivalence studies may not follow a normal distribution, the official approach to address this issue is to apply a natural logarithm transformation to the data before conducting statistical analyses [5,6]. Thus, regardless of the distribution of the original data, ln-transformation is always applied in the field of bioequivalence before statistical analysis [4–6]. After the analysis is completed, the results are transformed back to the original scale to interpret the findings in a clinically meaningful way, as we did in our study. The statistical analysis was performed using the typical bioequivalence criteria (90% confidence interval), and a decision regarding equivalence or not was made if the 90% confidence interval fell within the acceptance limits of 80.00–125.00% [5,6]. The statistical assessment followed the principles of equivalence testing, namely, the two-one-sided test (TOST) procedure [5,6]. The above-mentioned procedure was repeated several times (500) to allow for reliable estimates.

3. Results

Initially, an analysis was carried out to ascertain the optimal activation function for the hidden layers. The effectiveness of the linear and “softplus” activation functions was assessed. The implementation of the “softplus” activation function yielded faster and superior convergence. The “softplus” activation function demonstrated a convergence rate of 92%, while the linear activation function exhibited a convergence rate of 74%. Furthermore, the loss function’s final value was thrice higher when employing the linear activation function in contrast to the “softplus” activation function.

The next step of the analysis was the assessment of the activation function employed in the output layer. The “softplus” and linear functions were evaluated as potential activation functions. The aim of the study was to assess the degree of similarity between the generated data and the bell-shaped distribution of the source data. The outcomes are depicted in Figure 3.

As illustrated in Figure 3, the data generated for both R and T groups exhibited a bell-shaped distribution when a linear activation function was used for the output layer (Figure 3b). It was observed that the utilization of a linear activation function for the output layer resulted in a more optimal distribution as opposed to the “softplus” activation function, which led to a right-skewed distribution (Figure 3a). Overall, it appears that the utilization of the “softplus” activation function in the output layer yields an exponential-

like distribution for the resultant data, whereas the adoption of a linear activation function generates data with a bell-shaped distribution.

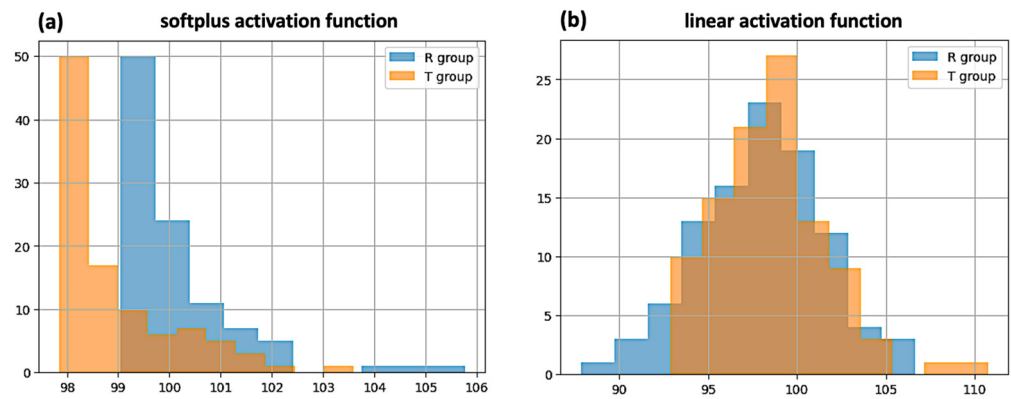


Figure 3. Distribution of the generated data for both R and T groups using the “softplus” (a) and linear (b) activation functions for the output layer.

The process of optimizing the number of epochs for model training was also executed. The values of 100, 500, 1000, 5000, and 10,000 underwent testing. Each of the variables mentioned above was used to train a VAE model and, after that, the trained model was used to generate data for the R and T groups. The results are illustrated in Figure 4.

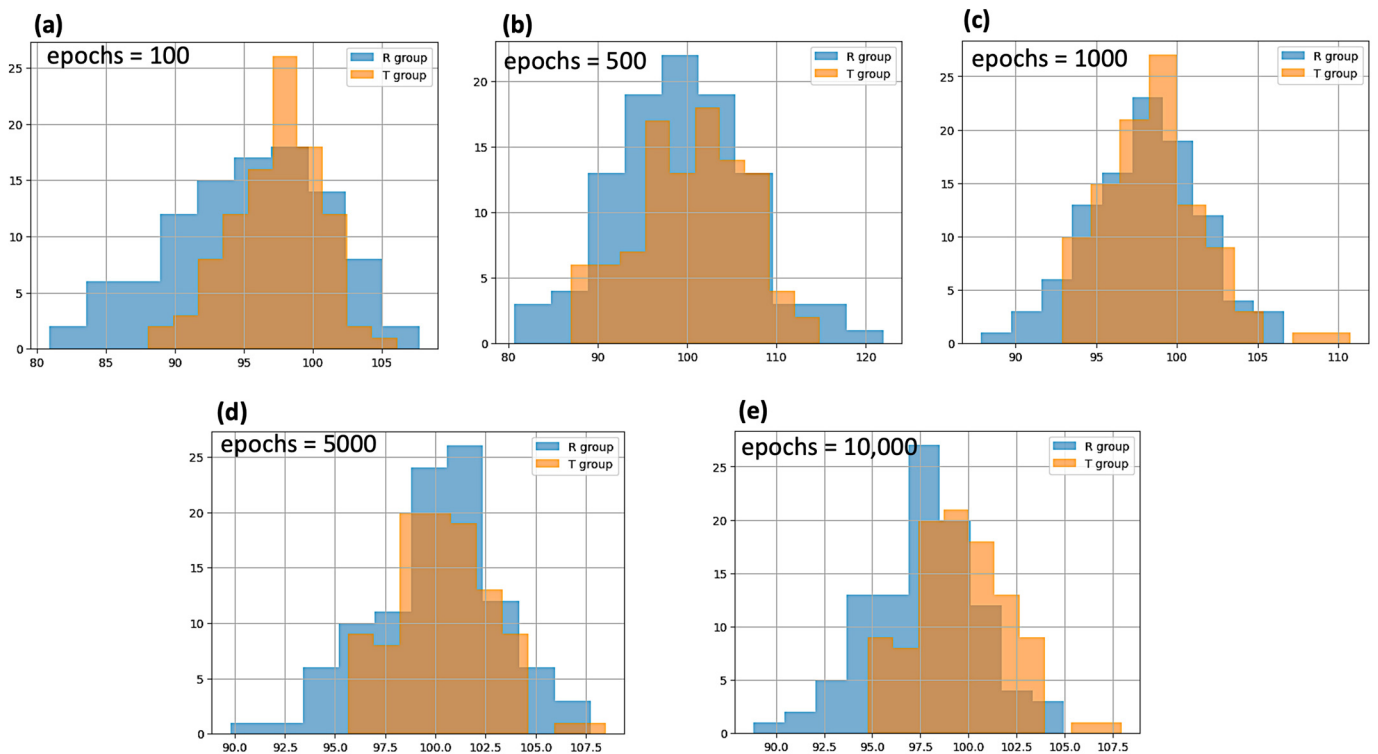


Figure 4. Distribution of the generated data for both the R and T groups using variational autoencoders with 100 (a), 500 (b), 1000 (c), 5000 (d), and 10,000 (e) epochs.

Figure 4 presents a graphical representation that reveals a considerable degree of variance in the data over 100 epochs (Figure 4a). Moreover, the data were not centered around the actual mean of both the R and T groups, which was 100. Similarly, this pertained to the situation in which 500 epochs were employed (Figure 4b). On the other hand, it was observed that increasing the number of epochs beyond 1000 made a significant difference in how well data were centered around the mean (which was 100) and how well variance

was placed within the desired range (Figure 4c). A similar pattern was further observed when the number of epochs was increased to 5000 and 10,000 (Figure 4d,e, respectively).

Finally, an investigation was carried out to ascertain the optimal number of hidden layers for both the encoder and decoder components. The investigation analyzed the values of 2, 3, and 4 as the number of hidden layers in both the encoder and decoder. As a consequence, overall numbers of 4, 6, and 8 hidden layers were evaluated, correspondingly. The shape and statistical properties of the generated data were evaluated for both the R and T groups in all cases (Figure 5).

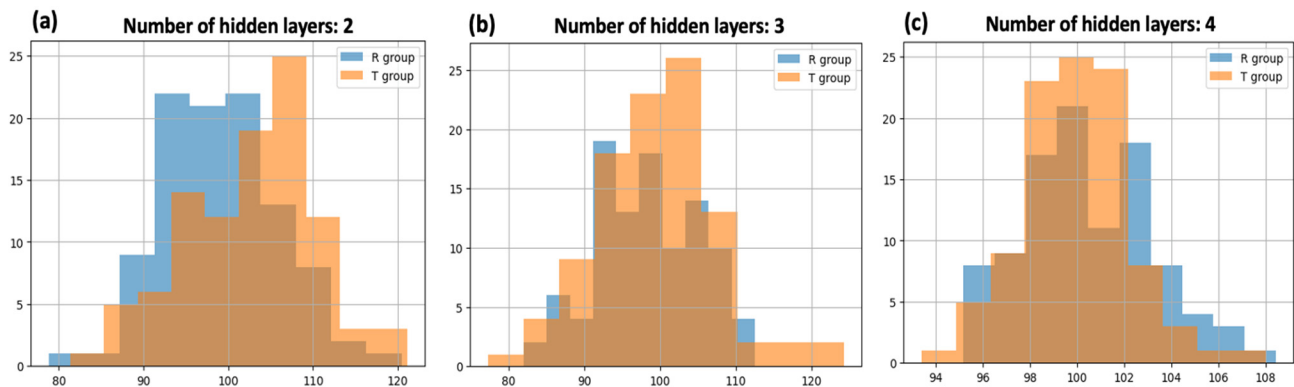


Figure 5. Distribution of the generated data for both the R and T groups using variational autoencoders with 2 (a), 3 (b), and 4 (c) hidden layers for the encoder and the decoder.

Figure 5 shows that the generated data did not exactly match the normal distribution of the original data and had a central tendency of about 100, especially when the encoder and decoder were limited to two hidden layers (Figure 5a). The findings indicate that there was a noteworthy enhancement in outcomes when the encoder and decoder were equipped with three hidden layers (Figure 5b). This particular arrangement facilitated a more effective capture of the shape and mean of the original dataset. This held true in cases where both the encoder and decoder had four hidden layers (Figure 5c).

In the next step, a statistical analysis was conducted to compare the properties of the generated datasets, which were obtained from subsamples of different sizes ranging from 10 to 90, with those of the original dataset. The assessment of equivalence (or non-equivalence) was explored for this objective. The research included exploration with diverse autoencoder configurations and data variability. Specifically, coefficient of variation values of 10%, 20%, and 40% were employed. Furthermore, the investigation examined the effects of distinct activation functions, specifically “softplus” and linear, on the hidden layer of the convolutional neural networks. The aforementioned procedure was implemented multiple times utilizing Monte Carlo simulations and the percentage of equivalence acceptance (i.e., the probability to reject the null hypothesis) was counted.

Figure 6 illustrates the probability of accepting equivalence under the TOST hypothesis. The diagram depicts the diverse subsample levels, alongside the two groups characterized by the three coefficients of variation values (10%, 20%, and 40%) and the two discrete activation functions employed for the hidden layers.

Figure 6 demonstrates the trend of equivalence acceptance as the subsample size proportion increased (from left to right) in the case of CVs 10% and 20% (Figure 6a,b, respectively). On the contrary, when the CV was equal to 40%, the probability of accepting equivalence rose with an increase in the subsample size (Figure 6c). This attribute could be observed in both the R and T groups. Ultimately, it seems that the probability of rejecting the null hypothesis (namely, declaring equivalence) is higher for the “softplus” when compared to the linear activation function.

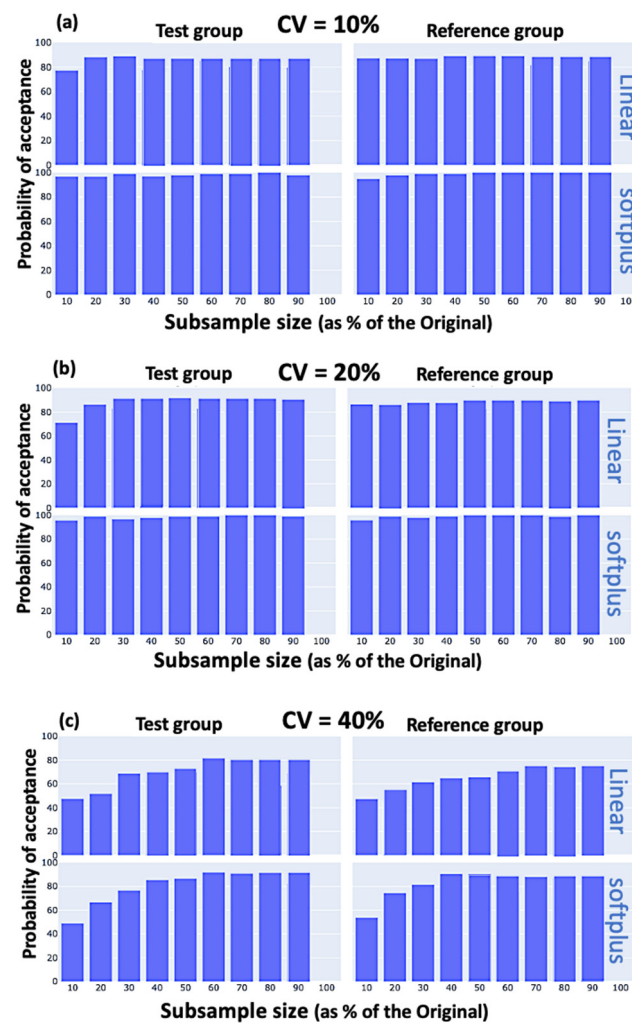


Figure 6. Probability of accepting equivalence between the original and the generated datasets for three levels of variability (CV): (a) 10%, (b) 20%, and (c) 40%. The results are shown separately for the test and reference groups, as well as the two types of activation functions (“softplus” and linear) used for the hidden layers.

In a subsequent step, T–R group comparisons were carried out. The datasets of the reference group, including the original, generated, and subsampled datasets, were compared to their corresponding counterparts in the test group (Figure 7). Several subsample sizes were assessed, spanning from 10% to 90%, with intervals of 10%. This statistical analysis aimed to investigate equivalence and was performed across multiple CV levels. It should be mentioned that these comparisons were exclusively carried out in the case in which the activation function of the hidden layers was “softplus”.

Figure 7 reveals that for low/medium CV values (10% and 20%), the probability of showing equivalence was quite high for all datasets. Especially for the original data (Figure 7a), where the probability is 100% since both the T and R groups were assumed to exhibit identical average performances at the endpoint. In the case of the subsampled (Figure 7b) and generated (Figure 7c) datasets, the probability of accepting equivalence was found to be low only in the cases using a very small part of the original dataset (e.g., 10% or 20%), and it increased to almost 100% acceptance when portions larger than 30% of this original sample size were used. When the high variability of the data was considered (CV = 40%), the observed performance was as expected. For the original dataset, the probability of acceptance fell to low values (close to 20%, Figure 7a). Similarly, the subsampled dataset showed poor performance since, not only for small portions of the original data but also for large parts, the probability of acceptance remained quite low

(Figure 7b). On the contrary, the VAE-generated (Figure 7c) dataset showed superior performance since, for all portions, the probability of acceptance was much higher than for the original and subsampled datasets. It is worth mentioning that even for low proportions (e.g., 10% or 20%), the statistical power of the VAE data was three times higher than that of the original data. For larger proportions, the probability of acceptance of the VAE-generated data reached rather high values (close to 80%), which was around four times higher than that of the original data.

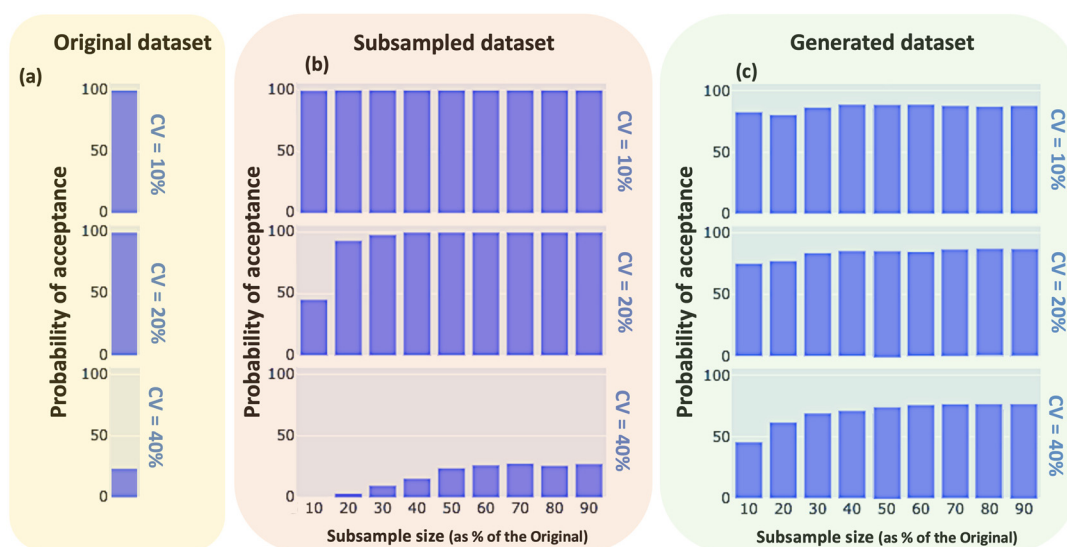


Figure 7. Probability of accepting equivalence between the test and reference groups for the original (a), subsampled (b), and generated (c) datasets. Three levels of variability (coefficient of variation, CV) were used: 10%, 20%, and 40%. In all cases, the “softplus” activation was used for the hidden layers, while both the test and reference groups were assumed to exhibit identical average performances.

In Figure 7, it becomes evident that for highly variable data, the application of the VAE worked rather efficiently as a data augmentation method. In all cases, the average performance (i.e., the mean endpoint value) between the two compared groups (T vs. R) was considered to be identical. In order to investigate additional situations where the two compared groups differed, Figure 8 was constructed. In Figure 8, the mean values for the T group are 100, 110, 125, and 150, while the mean value for the R group is always 100. This means that the T group is thought to be the same ($T/R = 1$) or different by 10%, 25%, or 50% ($T/R = 1, 1.1, 1.25, \text{ and } 1.50$). Also, the statistical characteristics of the original, subsampled, and generated data were investigated for a range of subsample sizes spanning from 10% to 90%. The impact of the CV on the clinical study outcomes was investigated based on two distinct values, specifically, 10% and 20%. Several iterations were used for each case, and the probability of accepting equivalence was subsequently calculated.

In all cases in Figure 8, a decrease in statistical power can be observed with an increase in the T/R ratio. In particular, for the original data (Figure 8a), a high probability of acceptance (almost 100%) can be observed when the two groups (T vs. R) do not differ ($T/R = 1$) or differ a little ($T/R = 1.1$). As the discrepancy between T and R gets larger (to 25% or 50%), a dramatic decrease in statistical power is observed. A similar performance is observed for the subsampled data (Figure 8b). Again, at low T/R ratios (e.g., 1 or 1.1), there is a high probability of equivalence acceptance, whereas, as the discrepancy between T and R becomes higher, the statistical power decreases and reaches almost zero values. Also, when small portions of the original data are used (around 10% to 30%), the probability of acceptance is even lower. For the VAE-generated datasets (Figure 8c), the results obtained for low T/R ratios (1 or 1.1) show a profile similar to the one observed before for the original and subsampled data. However, a desired performance with much higher statistical power can be observed when the two groups differ by 25%. For larger

discrepancies, namely, when $T/R = 1.50$, as expected, the probability of acceptance falls to zero since these high discrepancies in the average performance are outside the acceptance limits of equivalence. Since the acceptance limits in equivalence trials are between 80.00 and 125.00, there should be almost no probability of acceptance for discrepancies higher than 25%, namely, those exceeding the upper limit of 125.00 (or being below the lower limit of 80.00). This is reflected in Figure 8c, where the probability of acceptance is found to be almost zero when the T/R ratio is 1.25. The latter is in full agreement with the theoretical expectations. In other words, by applying the VAE-generated methodology, high statistical power is achieved for any variability level of the data, and, in addition, no false positives are observed when the discrepancy between the two groups exceeds the acceptance limits.

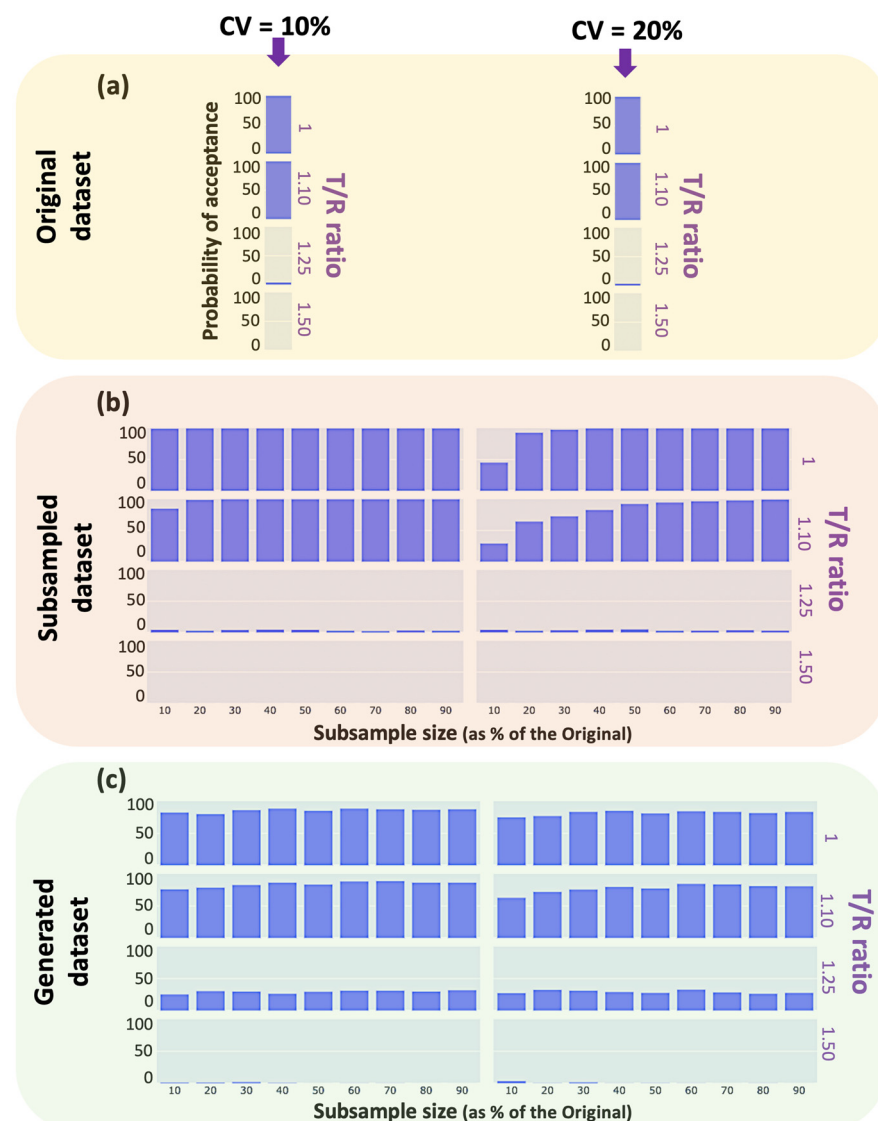


Figure 8. Probability of accepting equivalence between the test and reference groups for several ratios (1, 1.10, 1.25, 1.50) of the average test (T)/reference (R) performance. The comparisons were made separately for the original (a), subsampled (b), and generated datasets by the variational autoencoder (c). In all cases, the “softplus” activation function was used for the hidden layers and two levels of variability (coefficient of variation, CV) were used: 10% and 20%.

4. Discussion

The objective of the present study was to examine the process of reducing the required sample size of a clinical study by generating novel data through the utilization of a variational autoencoder [11]. Conducting a clinical trial with a large sample size can be extremely

expensive and time-consuming [1–4]. The recruitment, enrollment, and follow-up of a large number of participants require significant financial and logistical resources. Also, in some cases, using a large sample size might expose a larger number of participants to potential risks associated with the experimental intervention. The enrollment of many participants in a clinical trial can extend its duration, thus increasing the chances of there being external factors that influence the results and potentially affecting the study's validity. In addition, with a large number of participants, there is a higher likelihood of attrition and dropout during the course of the trial [1–4]. In this context, over the last few years, efforts have been made to conduct *in silico* clinical trials and/or virtually increase the sample size in order to face off all the previously mentioned drawbacks. Data augmentation techniques, such as the introduced VAE method, involve generating additional data points based on existing data. These methods can increase the effective sample size and improve the robustness of the analysis.

In order to accomplish this task, datasets were generated for two groups of volunteers (test vs. reference) using a normal distribution, while the conditions of an equivalence clinical trial were simulated. The most important factors affecting the outcome of a clinical trial refer to the mean difference between the two interventions (i.e., the relationships between the average T and R values), the variability of the measured endpoint (i.e., the coefficient of variation), and the sample size (i.e., the proportion of the “subsampled” population with regard to the original dataset). The impact of all these factors on the efficiency and robustness of the VAE methods was explored using a wide range of values. The desired situation would be one where the performance of the generated datasets is better than the one observed from the subsampled population. In other words, it would be desirable for the generated dataset to show equivalence when this truly exists (namely, when it is proven from the analysis of the original dataset) and to show non-equivalence when this is also shown from the original dataset. Any discrepancy with the original dataset indicates a deficiency in the analysis dataset. In order to investigate the efficiency of the VAE method, the performance of the VAE-generated datasets was compared with that of the original as well as with that of the subsampled dataset. In all cases, the performance of the VAE-generated data was found to be superior to any subsampled group and similar to the original (large) dataset (Figures 6–8). It is noteworthy that for high variabilities (Figure 7c), the performance of the VAE method became even better than with the original dataset.

The R group was assumed to have a mean endpoint value of 100, while several endpoint means were utilized for the T group; the T means were set at 100, 110, 125, and 150 in order to express identical performance (i.e., 100%) and a 10%, 25%, and 50% discrepancy, respectively. The aforementioned data were produced using varying coefficients of variation, specifically, 10%, 20%, and 40%. Subsequently, the data generated through a random process underwent subsampling at various percentages ranging from 10% to 90% of the original size. The subsampled data were then utilized to train a VAE, which was ultimately employed to generate novel data.

An initial analysis was conducted to determine the most suitable activation function for the hidden layers. The utilization of the “softplus” activation function resulted in more rapid and superior convergence. The initial stage involved conducting a test to determine the superior activation function between “softplus” and linear for the hidden layers. The findings indicate that the utilization of “softplus” activation leads to a higher rate of convergence in the neural network. Additionally, the average value of the loss function was observed to be three times greater when linear activation was employed as opposed to “softplus” activation. The “softplus” activation function is a modified version of rectified linear unit (ReLU) non-linearity designed to provide a continuous and differentiable approximation of the ReLU function. Its primary application is to ensure that the output of a computational model is always positive, thereby constraining the model predictions to a specific range [11].

The next step involved conducting a test to explore the performance of the two tested activation functions for the output layer (linear and “softplus”). In this case, the utilization of

the linear activation function demonstrated a tendency for the produced data to exhibit a greater degree of centralization around the true mean of the source datasets. Furthermore, it was observed that the choice of a linear activation function was more effective in representing the distribution shape of the initial data for both the R and T groups (Figure 3). Subsequently, a determination had to be made regarding the appropriate number of epochs used in the VAE. As illustrated in Figure 4, the most favorable number of epochs was 1000. This was because, for the cases where the number of epochs was 100 and 500, the generated data did not exhibit a satisfactory level of centralization around the true mean of 100. Furthermore, while the results for the cases where the number of epochs was 5000 and 10,000 differed slightly from the case with 1000 epochs, the differences were not significant. Based on the training time of the model, namely, the significantly more time needed as the number of epochs increased, it was concluded that utilizing 1000 epochs was the most effective option.

In addition, various numbers for the hidden layers were explored for both the encoder and the decoder. In this study, the general idea of deciding on architectural choices was based on the simplicity of the neural network and a trial-and-error procedure. Using a backward propagation procedure optimized the biases and weights of the network, which were reflected in the reduction of cost function values. For example, when utilizing two hidden layers for each, the original data bell shape was effectively represented. However, the generated data exhibited poor centering around the true mean (i.e., 100 for the R group). In instances where there were three or four hidden layers, the dataset typically consisted of approximately 100 observations and exhibited a well-defined bell shape. In accordance with Occam's razor principle, it is recommended that the optimal number of hidden layers for both the encoder and decoder be set at three (Figure 5).

After adjusting the VAE system's hyperparameters (number of hidden layers, activation functions, number of neurons, etc.), simulation results showed that the VAE system could successfully recreate data that were similar to those of the original dataset. For all scenarios studied, it was shown that the subsampled datasets, as expected, had the worst performance. Because of the reduced sample size, the subsampled datasets failed to imitate the behavior of the original data since they did not exhibit the required statistical power to show equivalence whenever this existed (Figures 7b and 8b). On the contrary, the generated data using the VAE showed increased statistical power when the data exhibited either low or high variability (Figures 7c and 8c). It should be underlined that for all low variabilities, the VAE-generated data exhibited a performance similar to the one observed with the original data, even when only a small portion of the original data was used. When the variability of the original data was low (i.e., CV = 10% or 20%), the use of one-third of the original sample through the VAE system could lead to statistical power that was only slightly less than that observed with the original data (Figure 7c). By increasing the proportion of subjects to around 50–70%, almost the same statistical power could be achieved. This attribute became even more evident in the case of highly variable data. In this situation, the VAE system not only succeeded in showing a similar performance as that for the original data but also presented even better behavior (Figure 7c). For highly variable data, the VAE could act as a noise (variability) filter and lead to increased statistical power.

It should be stated that high variability ("noise") is an issue of paramount importance in the field of bioequivalence [4]. Highly variable drugs refer to medicines that exhibit substantial variability in their pharmacokinetic parameters, specifically their absorption, distribution, metabolism, and elimination processes, when administered to individuals. However, for highly variable drugs, achieving strict bioequivalence can be challenging due to the inherent variability in their pharmacokinetics. When conducting bioequivalence studies for such drugs, the variability between individuals' responses can lead to wider confidence intervals, making it more difficult to demonstrate equivalence within the standard regulatory requirements. As a result, regulatory authorities often have specific guidelines and acceptance criteria for highly variable drugs to account for the expected variability. These criteria could include widening the acceptable range for certain pharmacokinetic

parameters (e.g., C_{max}), using different statistical methods to evaluate bioequivalence (e.g., scaled equivalence), or increasing the number of samples in the study [4]. Thus, it is crucial to find methods to decrease this unwanted variability and avoid increasing the number of study participants, the costs, and the complexity of the study.

The good performance of the VAE-generated data was also shown in cases where the two comparison groups differed (Figure 8). Again, it was shown that the use of a VAE can mimic the performance of the original data and, in cases of high variability, result in higher statistical power (Figure 8c). These results are consistent with the existing literature in the field of highly variable drugs [4]. The findings indicate that a sufficient level of acceptable probability can be achieved even for data with high variability by utilizing only 40% of the original data for accepting equivalence. Conversely, for data with low variability, a very small proportion (i.e., 10%) of the original data can be adequate (Figure 6).

The present investigation is a novel attempt to employ an artificial intelligence technique in the field of clinical trials, aiming at reducing sample size. As mentioned above, the high variability of data encountered in bioequivalence studies is a limiting factor for proving equivalence and leads to inflation in type II errors. There are two ways to tackle this problem: increase the sample size, which is followed by increases in the duration of the study, cost, complexity, and ethical concerns [5,6], or apply computational approaches (such as the use of scaled limits) [4–6]. In this study, neural networks were introduced as a tool to increase the statistical power of clinical studies without recruiting additional volunteers and avoiding all the drawbacks that follow. Thus, the use of VAE in the analysis of clinical trials can reduce the time needed for completing a trial, reduce costs, and certainly be fully in line with the ethical concerns of reducing unnecessary human exposure in clinical trials.

In particular, this study introduced the use of variational autoencoders in the statistical analysis of clinical trials. A literature search revealed that there is a lack of research pertaining to the use of artificial intelligence in clinical data augmentation. It is evident that while there are articles that evaluate data augmentation methods for image data, there is a dearth of literature that comprehensively assesses data augmentation techniques for numerical data in the clinical trials field [17,18]. A recent study tried to increase the quantity of data derived from a limited sample size [19]. This study aimed to create a simulated population of human coronary arteries for the purpose of conducting *in silico* clinical trials on stent design. The findings exhibit promise; however, there is a lack of information pertaining to the requisite sample size needed to maintain the statistical properties of the produced data in comparison to the original dataset. One objective of this investigation was to examine precisely that [19].

The recognition of the potential of AI in clinical trials has prompted the examination and discourse of its diverse implementations in the literature [7,10]. Artificial intelligence possesses the ability to predict the probability of trial or site failure, along with patient outcomes. Also, AI possesses the ability to scrutinize medical records to ascertain suitable groups for clinical trials [10]. The acceleration of trial recruitment can be achieved through the implementation of AI technologies, which can effectively notify medical personnel and patients about available trial opportunities. Moreover, the implementation of AI technology facilitates the streamlined monitoring of clinical trials, while the simplification of entry criteria can potentially improve accessibility for prospective participants.

Thus far, data augmentation approaches have been applied in the field of image analysis. Indeed, considerable emphasis has been placed on the utilization of data augmentation techniques in the field of computer vision, particularly with respect to images. Quite recently, Goceri presented a comprehensive survey of prior studies that have investigated the field of data augmentation in the context of medical imagery [17]. The objective of these studies was to enhance image data through the implementation of various techniques such as random rotation, noise addition, sharpening, and shifting [20,21]. Other advanced techniques, including generative adversarial networks, have also been proposed for similar purposes [22,23]. In the same vein, researchers have attempted to generate authentic fetal ultrasound images in the context of medical imaging [24]. The process involved the

extraction of the region of interest from each image using various techniques, followed by the application of a de-noising algorithm and, ultimately, an approximation method. In the present study, as the data consisted of randomly generated numerical values conforming to a normal distribution, the initial two steps were deemed unnecessary. Variational autoencoders have also been used in the augmentation of medical images [25]. Additionally, Chadebec and colleagues attempted to implement data augmentation techniques for high-dimensional data, such as images [26].

VAEs have already been successfully applied in diverse scientific areas. In a study, a conditional VAE was used for intrusion detection in an internet of things network [27]. The significance of the latter is quite important as the economic importance of intrusion detection systems expands, rendering them susceptible to future intrusion attacks. The authors utilized a conditional VAE with a distinct architecture that incorporated intrusion labels within the decoder layers. The proposed methodology exhibited several advantages, such as a lower level of complexity, high accuracy, and superior classification outcomes when compared to other well-known classifiers [27]. Another study utilized VAE in the area of choice modeling [28]. Choice modeling plays a pivotal role in transportation research, particularly in the areas of demand forecasting and infrastructure development. The process comprises two primary stages: the generation of a set of choices and the modeling of the decision-making process based on the provided choice set. The procedure of generating choice sets is very important as the inclusion or exclusion of certain options in the choice set can lead to inaccuracies in estimating parameters. The authors developed a generalized extreme value model that served to connect the value-added evaluation approach with choice modeling [28]. In another recent study, the authors applied a VAE model along with image processing methods in game design [29]. This study is considered to be the first to investigate various mathematical properties associated with VAE models. The VAE model demonstrated its proficiency in data clustering, and it was observed to be particularly efficient in generating images that exhibit a certain graphical structure or in managing and creating images that have low resolution demands [29]. Finally, another study presented a novel approach that combined a disentangled VAE with a bidirectional long short-term memory network backend in order to detect anomalies in heart rate data collected during sleep using a wearable device [30]. The performance of this model was compared with that of other established anomaly detection algorithms. It was shown that the developed model exhibited superior performance across a wide range of scenarios and with all participants under consideration [30].

In this study, VAEs are proposed as a data augmentation method in clinical trials, namely, as a way to reduce the required sample size. VAEs can be considered an extension of traditional autoencoders [11]. In contrast to autoencoders, variational autoencoders employ an encoder network that maps input data to a multivariate normal distribution, rather than a fixed point. In other words, VAEs strive to create a correlation between the input data and a probability distribution throughout the latent space [11–13]. The use of VAEs offers several advantages; for example, they have been identified as a highly effective and valuable approach for the development of generative models, namely, models that are utilized to produce novel synthetic or artificial data based on existing data. A notable benefit of VAEs in comparison to conventional autoencoders lies in their ability to generate novel data from the same embedding distribution as the input data through sampling from the embedded distribution. These unique features were exploited in this study in order to virtually increase the sample size of a clinical trial. This study showed that by applying VAE, it became feasible to use only 20% of the original dataset without changing the true outcome of the study (Figure 6a,b). It is worth noting that even for data exhibiting high variability (e.g., 40%), the use of a VAE can reduce the need for a sample size to less than half of the one typically needed for a trial (Figure 6c). This dramatic reduction in sample size can accelerate clinical trials, significantly reduce costs, and certainly diminish human exposure. To the best of our knowledge, this study represents the first attempt

to employ autoencoders (and neural networks in a general sense) within the domain of clinical research aimed at reducing the necessary sample size.

A limitation of this study is the limited number of iterations performed for each scenario. Considering the quite long duration of execution, the completion of 500 runs consumed a considerable amount of time. However, within the realm of clinical trials, this quantity may be deemed modest, and, in any case, this number of iterations offered an adequate degree of convergence, as depicted in Figures 6–8. Another possible limitation is the fact that the statistical analysis performed in this study was based only on the equivalence criterion, expressed in the context of the two-one-sided t-test procedure [4]. However, this test is frequently utilized in clinical trials, required by regulatory authorities (e.g., FDA, EMA), and explained in detail from a regulatory standpoint [5,6]. Additional clinical designs (e.g., crossover, replicate, etc.) and statistical hypotheses must be explored, as in the case of non-inferior and superior clinical trials. In addition, with regard to the model architecture, it may be worthwhile to investigate more sophisticated models that incorporate additional hidden layers and/or a greater number of neurons per layer, while taking computational efficiency into account. Also, it is worth exploring additional activation functions that could be applied to the hidden and output layers. Finally, application to real clinical data, as in the case of any computational approach, is necessary before adopting it in practice. It should be underlined that the integration of real data into simulated studies is essential for improving the accuracy, reliability, and applicability of the simulations, making them valuable tools for understanding real-world phenomena, making predictions, and informing decision-making processes. A comparative analysis of applying the introduced VAE procedure to real clinical data would allow for the identification of its applicability and efficiency.

5. Conclusions

The aim of this study was to introduce the use of neural networks, particularly variational autoencoders, as a tool to virtually increase the sample size in clinical studies and thereby decrease the required number of actual participants. This study began by developing the most appropriate architecture for the VAE and tuning hyperparameters, such as the number of hidden layers (for both the encoder and the decoder), number of neurons per layer, selection of the activation function, number of epochs, and weights. The next step involved applying the developed VAE model in simulated Monte Carlo clinical studies under various scenarios that can occur in practice. These scenarios included several levels of variability in the measured endpoints, different average performances between compared groups, and varying sizes of the subsampled group. The efficiency of the VAE-generated data was then compared with that of the original data and also against that of the subsampled data. In all cases, using the VAE-generated data resulted in an increase in the statistical power of the study, especially in cases of high variability. Importantly, the type I error was kept at low values and remained at the same level as with the original data, while the type II error of the VAE method was even lower compared to the original datasets. Overall, the combined use of VAE with Monte Carlo simulated clinical trials demonstrated the desired performance, leading to less human exposure in clinical studies and significantly reduced costs and time for trial completion. To the best of our current understanding, this study represents a novel effort to employ autoencoders and neural networks within the realm of clinical research, specifically aiming to reduce the necessary sample size.

Author Contributions: Conceptualization, V.D.K.; methodology, V.D.K.; software, D.P.; validation, D.P.; formal analysis, D.P.; investigation, D.P.; resources, V.D.K.; data curation, D.P.; writing—original draft preparation, D.P.; writing—review and editing, V.D.K.; visualization, D.P. 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: Not applicable.

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

References

1. Sakpal, T.V. Sample size estimation in clinical trial. *Perspect. Clin. Res.* **2010**, *1*, 67–69.
2. Wang, X.; Ji, X. Sample Size Estimation in Clinical Research: From Randomized Controlled Trials to Observational Studies. *Chest* **2020**, *158*, S12–S20. [[CrossRef](#)] [[PubMed](#)]
3. Malone, H.E.; Nicholl, H.; Coyne, I. Fundamentals of estimating sample size. *Nurse Res.* **2016**, *23*, 21–25. [[CrossRef](#)] [[PubMed](#)]
4. Karalis, V. Modeling and Simulation in Bioequivalence. In *Modeling in Biopharmaceutics, Pharmacokinetics and Pharmacodynamics. Homogeneous and Heterogeneous Approaches*, 2nd ed.; Iliadis, A., Macheras, P., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 227–255.
5. European Medicines Agency; Committee for Medicinal Products for Human Use (CHMP). *Guideline on the Investigation of Bioequivalence*; CPMP/EWP/QWP/1401/98 Rev. 1/Corr**; Committee for Medicinal Products for Human Use (CHMP): London, UK, 20 January 2010. Available online: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf (accessed on 29 May 2023).
6. Food and Drug Administration (FDA). Guidance for Industry. Bioavailability and Bioequivalence Studies Submitted in NDAs or INDs—General Considerations. Draft Guidance. U.S. Department of Health and Human Services Food and Drug Administration. Center for Drug Evaluation and Research (CDER). December 2013. Available online: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/bioavailability-and-bioequivalence-studies-submitted-ndas-or-ind-s-general-considerations> (accessed on 29 May 2023).
7. Askin, S.; Burkhalter, D.; Calado, G.; El Dakrouni, S. Artificial Intelligence Applied to clinical trials: Opportunities and challenges. *Health Technol.* **2023**, *13*, 203–213. [[CrossRef](#)] [[PubMed](#)]
8. Harrer, S.; Shah, P.; Antony, B.; Hu, J. Artificial Intelligence for Clinical Trial Design. *Trends Pharmacol. Sci.* **2019**, *40*, 577–591. [[CrossRef](#)]
9. Delso, G.; Cirillo, D.; Kaggie, J.D.; Valencia, A.; Metser, U.; Veit-Haibach, P. How to Design AI-Driven Clinical Trials in Nuclear Medicine. *Semin. Nucl. Med.* **2021**, *51*, 112–119. [[CrossRef](#)]
10. The Alan Turing Institute. Statistical Machine Learning for Randomised Clinical Trials (MRC CTU). Available online: <https://www.turing.ac.uk/research/research-projects/statistical-machine-learning-randomised-clinical-trials-mrc-ctu> (accessed on 29 May 2023).
11. Chollet, F. *Deep Learning with Python*, 2nd ed.; Manning; Simon and Schuster: New York, NY, USA, 2021.
12. Atienza, R. *Advanced Deep Learning with Keras: Apply Deep Learning Techniques, Autoencoders, GANs, Variational Autoencoders, Deep Reinforcement Learning, Policy Gradients, and More*; Packt Publishing: Birmingham, UK, 2018.
13. Kingma, D.; Welling, M. *An Introduction to Variational Autoencoders (Foundations and Trends(r) in Machine Learning)*; Now Publishers Inc.: Hanover, MA, USA, 2019.
14. Henderson, H. *Artificial Intelligence: Mirrors for the Mind (Milestones in Discovery and Invention)*, 1st ed.; Chelsea House Pub: Singapore, 2007; 176p.
15. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4th ed.; Pearson: London, UK, 2021; 1136p.
16. Yang, Y.; Ye, Z.; Su, Y.; Zhao, Q.; Li, X.; Ouyang, D. Deep learning for in vitro prediction of pharmaceutical formulations. *Acta Pharm. Sin. B* **2019**, *9*, 177–185. [[CrossRef](#)]
17. Goceri, E. Medical image data augmentation: Techniques, comparisons and interpretations. *Artif. Intell. Rev.* **2023**, 1–45. [[CrossRef](#)]
18. Kebaili, A.; Lapuyade-Lahorgue, J.; Ruan, S. Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review. *J. Imaging* **2023**, *9*, 81. [[CrossRef](#)]
19. Pleouras, D.; Sakellarios, A.; Rigas, G.; Karanasiou, G.S.; Tsompou, P.; Karanasiou, G.; Kigka, V.; Kyriakidis, S.; Pezoulas, V.; Gois, G.; et al. A Novel Approach to Generate a Virtual Population of Human Coronary Arteries for In Silico Clinical Trials of Stent Design. *IEEE Open J. Eng. Med. Biol.* **2021**, *20*, 201–209. [[CrossRef](#)]
20. Khan, A.R.; Khan, S.; Harouni, M.; Abbasi, R.; Iqbal, S.; Mehmood, Z. Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification. *Microsc. Res. Tech.* **2021**, *84*, 1389–1399. [[CrossRef](#)]
21. Maqsood, S.; Damaševičius, R.; Maskeliūnas, R. Hemorrhage Detection Based on 3D CNN Deep Learning Framework and Feature Fusion for Evaluating Retinal Abnormality in Diabetic Patients. *Sensors* **2021**, *21*, 3865. [[CrossRef](#)]
22. Chen, X.; Lian, C.; Wang, L.; Deng, H.; Kuang, T.; Fung, S.H.; Gateno, J.; Shen, D.; Xia, J.J.; Yap, P.T. Diverse data augmentation for learning image segmentation with cross-modality annotations. *Med. Image Anal.* **2021**, *71*, 102060. [[CrossRef](#)] [[PubMed](#)]
23. Barile, B.; Marzullo, A.; Stamile, C.; Durand-Dubief, F.; Sappey-Marini, D. Data augmentation using generative adversarial neural networks on brain structural connectivity in multiple sclerosis. *Comput. Methods Programs Biomed.* **2021**, *206*, 106113. [[CrossRef](#)] [[PubMed](#)]

24. Athalye, C.; Arnaout, R. Domain-guided data augmentation for deep learning on medical imaging. *PLoS ONE* **2023**, *18*, e0282532. [[CrossRef](#)] [[PubMed](#)]
25. Pesteie, M.; Abolmaesumi, P.; Rohling, R.N. Adaptive Augmentation of Medical Data Using Independently Conditional Variational Auto-Encoders. *IEEE Trans. Med. Imaging* **2019**, *38*, 2807–2820. [[CrossRef](#)]
26. Chadebec, C.; Thibeau-Sutre, E.; Burgos, N.; Allasonniere, S. Data Augmentation in High Dimensional Low Sample Size Setting Using a Geometry-Based Variational Autoencoder. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 2879–2896. [[CrossRef](#)]
27. Lopez-Martin, M.; Carro, B.; Sanchez-Esguevillas, A.; Lloret, J. Conditional Variational Autoencoder for Prediction and Feature Recovery Applied to Intrusion Detection in IoT. *Sensors* **2017**, *17*, 1967. [[CrossRef](#)]
28. Yao, R.; Bekhor, S. A Variational Autoencoder Approach for Choice Set Generation and Implicit Perception of Alternatives in Choice Modeling. *Transp. Res. Part B Methodol.* **2022**, *158*, 273–294. [[CrossRef](#)]
29. Mak, H.W.L.; Han, R.; Yin, H.H.F. Application of Variational AutoEncoder (VAE) Model and Image Processing Approaches in Game Design. *Sensors* **2023**, *23*, 3457. [[CrossRef](#)]
30. Staffini, A.; Svensson, T.; Chung, U.; Svensson, A.K. A Disentangled VAE-BiLSTM Model for Heart Rate Anomaly Detection. *Bioengineering* **2023**, *10*, 683. [[CrossRef](#)] [[PubMed](#)]

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.